Universidade Federal de Campina Grande

Centro de Engenharia Elétrica e Informática

Coordenação de Pós-Graduação em Ciência da Computação

Modelos Baseados em PPM para Previsão de Trajetórias Utilizando Informações Contextuais

Francisco Dantas Nobre Neto

Tese submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Ciência da Computação

Linha de Pesquisa: Sistemas de Informação Geográfica

Cláudio de Souza Baptista, Ph.D.
(Orientador)

Cláudio Elizio Calazans Campelo, Ph.D.
(Orientador)

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

N754m Nobre Neto, Francisco Dantas.

Modelos baseados em PPM para previsão de trajetórias utilizando informações contextuais / Francisco Dantas Nobre Neto.– Campina Grande, 2017.

132 f. il. color.

Tese (Doutorado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2017.

"Orientação: Prof. Dr. Cláudio de Souza Baptista, Prof. Dr. Cláudio Elizio Calazans Campelo".

Referências.

1. Previsão de Trajetórias. 2. *Prediction By Partial Matching (PPM)*. 3. Reconhecimento de Padrões. 4. Aprendizagem de Máquina. I. Baptista, Cláudio de Souza. II. Campelo, Cláudio Elizio Calazans. III. Título.

CDU 004.5(043)

Resumo

Com a ampla difusão de *smartphones* equipados com dispositivos GPS (*Global Positioning System*), rastrear a localização de objetos (como pessoas e veículos) tem sido uma tarefa mais factível, resultando em novas oportunidades de pesquisas em variadas áreas do conhecimento. Dentre estas oportunidades, esta pesquisa lida com o desafio da área de previsão de rotas e destinos. Saber antecipadamente o destino de um usuário assim que ele inicia um deslocamento tem muitas utilidades práticas, tais como: indicar rotas menos congestionadas ou vias mais seguras, e sugerir a visita a algum ponto de interesse (POI) antes do destino almejado. Sistemas que fornecem previsão de rota e destino estão disponíveis comercialmente, no entanto, estes podem requerer interações constantes do usuário. Para deslocamentos diários, porém, a necessidade de uma interação frequente do usuário com um aplicativo pode tornar seu uso pouco prático e pouco ubíquo. Além disso, muitos trabalhos que apresentam modelos de previsão de rotas e destinos, disponíveis na literatura, não contemplam uma importante informação contextual, que é o papel que os lugares visitados representam para um usuário (por exemplo, se é sua *casa* ou seu local de *trabalho*). Não obstante, a maioria dos preditores disponíveis não possuem a funcionalidade de prever lugares nunca visitados. Esta tese de doutorado propõe uma família de métodos de predição baseada no algoritmo de compressão de dados *Prediction by Partial Matching* (PPM). Ainda com relação a esta pesquisa, é proposto um mecanismo capaz de identificar que uma rota em curso está sendo realizada pela primeira vez e, portanto, ter a possibilidade de prever um destino ainda não visitado. Neste estudo, também foram implementados outros preditores consolidados na literatura, que são as Cadeias de Markov e as Cadeias Ocultas de Markov, utilizados para comparação. É importante observar que ambos os preditores são capazes de prever apenas o destino de um trajeto, ao invés da rota restante. Nos experimentos realizados, foram utilizadas as métricas de *Precisão, Recall* e *Medida-F* (*F1 Score)*, com validação cruzada (contendo 10 partições mutuamente exclusivas), para avaliação dos modelos de previsão implementados. A base de dados utilizada nesta pesquisa é composta por mais de 1.500 rotas, coletadas por aproximadamente três meses, referentes a 21 usuários. Os preditores baseados em PPM apresentaram resultados competitivos (ou superiores) comparados aos da literatura.

Abstract

Thanks to the widely diffusion of *smartphones* with GPS devices natively embedded, the task of tracking object locations, such as people or vehicles, is more feasible nowadays, fostering new research opportunities. Among these new opportunities, this work addresses the challenge of route and destination prediction. Knowing in advance the destination where a user might reach as soon as he or she starts to move can be useful in various situations. For instance, to suggest to users less jammed or safer routes, as well to warn about points of interest located along their route. There are commercial systems capable of predicting destination and routes, however, these systems usually require frequent user interaction. Nonetheless, such a requirement could make the application unusable for daily routines. Moreover, most existing works do not consider an important contextual information: the information about the places that the users visit, i.e., the role that the places play to the user (for instance, if the place is *home* or *work*). In addition, most predictors described in the literature are not able to predict places that users have never visited. This thesis proposes a family of algorithms based on *Prediction by Partial Matching* (PPM). Furthermore, this work proposes a mechanism for identifying whether a route is being performed for the first time, resulting in the feasibility for predicting a never visited place. This research also provides a comparison between our proposed predictors, and the predictors based on *Markov Models* and *Hidden Markov Models* (HMM), which have been used in related works. It is important to mention that both Markov and HMM predictors that we implemented are able to predict just the destination, instead the remaining route. For the statistical assessment of the predictors, the metrics *Precision, Recall* and *F1 Score* are used, together with the process of 10-fold cross-validation. The database contains about 1,500 routes extracted from 21 users, gathered for three months. The predictors based on PPM performed similarly (or better) than others reported in the literature.

Dedico esta tese aos meus pais, Antônio Telmi Dantas Nobre e Soraya Formiga Mariz Dantas, e à minha esposa, Niara Fernandes Barbosa Formiga Dantas, pelo amor, pela paciência e por todo o apoio que vocês me dão sempre.

Agradecimentos

Agradeço a Deus, por estar comigo durante toda esta caminhada de quatro anos, sustentando a minha fé e abençoando meu caminho, principalmente nos momentos de maiores incertezas. Agradeço, ainda, por Ele ter me apresentado desafios e reflexões de ordem pessoal que foram pertinentes para que eu pudesse, acima de tudo, melhorar como ser humano. Graças a essa caminhada conjunta, eu pude ter a força e a disposição necessárias para superar alguns obstáculos.

Agradeço aos meus amados pais, Telmi Dantas e Soraya Dantas, e à minha amada esposa, Niara Formiga. Esta é uma vitória nossa! Aos meus pais, que sempre me proporcionaram as coisas mais importantes na vida, amor, apoio e carinho, fundamentais para enfrentar uma empreitada desafiadora como esta. A eles, eu agradeço, ainda, por não medirem esforços em investir na melhoria do meu conhecimento, sempre dizendo que este é o bem mais precioso que os pais podem deixar aos filhos. E de fato é!

À minha bela e amada esposa, Niara Formiga, meu agradecimento especial. Ela se dedicou muito para que este trabalho pudesse ter sido feito, me confortou e tivemos que abdicar, juntos, de muitos encontros com a família e de diversões. Você sempre é compreensiva e quero dizer que você teve, e tem, uma importância grande nesta conquista.

Aos meus familiares, em especial aos meus avós paternos, *dona* Maria Leite (Leca) e *seu* Francisco Dantas (*in memorian*), e aos meus avós maternos, *dona* Maria Guiomar e *seu* Lauro Mariz (*in memorian*), por estarem no meu convívio pessoal, por entenderem minha ausência em alguns momentos comemorativos e por darem palavras encorajadoras. É muito bom compartilhar das suas histórias e das suas experiências, isso me incentiva bastante a querer melhorar como pessoa. A família também sabe como a caminhada é longa, e fui muito bem confortado.

Aos meus amigos, pessoais e profissionais, todos vocês tiveram contribuições importantes. Sou muito grato pelas conversas descontraídas que tivemos durante este período. Sem nem saberem, vocês me ajudaram a aliviar vários momentos de tensão.

Agradeço bastante aos meus orientadores de doutorado, Professor Cláudio Baptista e Professor Cláudio Campelo. Vocês me orientaram a como se fazer pesquisa científica, algo que levarei para sempre. Sempre me deram espaço para pesquisar o que

eu julgava ser importante, sem qualquer imposição, mas com orientação. Tanto no campo profissional como no pessoal, vocês me orientaram com sugestões valiosas. Ainda, obrigado pelo entendimento dos momentos difíceis que tive que enfrentar. Vocês foram verdadeiros orientadores, e espero multiplicar o que vocês fizeram por mim, inclusive, dando retorno à sociedade, que tanto investe na Educação.

Agradeço aos profissionais do Programa de Pós-Graduação de Computação, da Universidade Federal de Campina Grande (UFCG), especialmente, à Professora Joseana Fechine e ao Professor Herman Gomes, pela disponibilidade, colaboração e atenção conosco neste trabalho. Vocês contribuíram de maneira ímpar, com colocações construtivas e de maneira muito serena. Ao Professor Herman, dedico um agradecimento, ainda, por ter me apresentado ao Professor Baptista para orientação e ter permitido mudança de orientação, de maneira muito amigável e conversada.

Agradeço pela imensa colaboração dos avaliadores externos, Professora Valéria Times e Professor José Macêdo, respectivamente, da Universidade Federal de Pernambuco (UFPE) e da Universidade Federal do Ceará (UFC). Os senhores têm muita propriedade no assunto desta tese, são reconhecidos na comunidade científica e, mesmo com restrições de distância e tempo, aceitaram, sem qualquer dificuldade, avaliar e colaborar com esta tese. Tenham certeza que vocês enriqueceram muito este trabalho.

Agradeço aos amigos do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), instituição onde trabalho, em especial àqueles do *Campus* Campina Grande, por terem me ajudado em muitas situações e entendido a correria deste momento, como ajustes de horários e liberação para me dedicar ao doutorado.

Agradeço a todos os meus Professores acadêmicos que tive, e tenho, que tentam passar o conhecimento da melhor forma possível. Em especial, ao Professor Leonardo Batista, da Universidade Federal da Paraíba (UFPB). Em um momento inicial e importante da minha vida acadêmica, fui contemplado para fazer parte da sua equipe de trabalho, o Programa de Educação Tutorial (PET) de Computação, da UFPB, e, desde então, venho tendo interesse crescente em pesquisar e estudar. Agradeço pela confiança, e tenha certeza que cresci profissionalmente e pessoalmente com esta participação.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

ANOVA	Análise de Variância
ATIS	*Advanced Traffic Information System*
DBSCAN	*Density-based Spatial Clustering of Applications with Noise*
GPS	*Global Positioning System*
HMM	*Hidden Markov Model*
ICP	*Increased Confidence Prediction*
ITS	*Intelligent Transportation System*
LBS	*Location-Based Services*
POI	*Point of Interest*
PPM	*Prediction by Partial Matching*
RC	Razão de Compressão
ST-DBSCAN	*Spatial-Temporal DBSCAN*

Capítulo 1

Introdução

Nos últimos anos, têm sido dedicados esforços significativos da Tecnologia da Informação (TI) no setor de transportes e mobilidade urbana (WOLFSON, SISTLA e XU, 2012). Sistemas que planejam rotas de deslocamentos para usuários e que dispõem de informação do trânsito em tempo real são alguns exemplos nos quais os serviços da TI podem facilitar o planejamento e a locomoção urbana (WOLFSON, SISTLA e XU, 2012) (VARRIALE, MA e WOLFSON, 2013). Tais sistemas estão inseridos no âmbito de uma temática conhecida como Sistema Inteligente de Transporte, ou ITS (*Intelligent Transportation System*).

ITS é uma área multidisciplinar, em que são envolvidos conhecimentos de computação e de engenharia de transportes. De maneira geral, trata-se do desenvolvimento de aplicativos com o intuito de prover serviços inovadores para melhoria no uso de transportes e no gerenciamento de tráfego. Nesta área, são associados sistemas de informação, comunicação e sensores com o ambiente de infraestrutura física existente nas cidades ou regiões, para melhoria, por exemplo, do tráfego de uma região (FEI, LUB e LIUC., 2011). Dentre os variados desafios relacionados à área de ITS, que podem ser enfrentados com o apoio dos serviços de TI, estão: a identificação de congestionamentos em determinadas vias; a informação sobre itinerários de ônibus; e a previsão de destino e rotas para um deslocamento em curso ou que ainda vai se iniciar.

De particular interesse neste trabalho, está a temática relacionada à previsão de destino e rotas. Saber antecipadamente qual vai ser o destino de um usuário que inicia seu deslocamento é útil em diversos contextos. De posse dessa informação e com dados do

tráfego obtidos em tempo real, um sistema computacional pode sugerir, por exemplo, rotas menos congestionadas, mais rápidas, seguras ou adequadas para quem realiza um passeio turístico (por exemplo, rotas cênicas) (BRILHANTE *et al*., 2015). Além disso, há também a possibilidade de sugestão de pontos de interesse (POI, do inglês *Point of Interest*), como, por exemplo, uma padaria ou farmácia localizada no percurso para o destino do usuário. A necessidade de previsão de rotas e destino tem sido alvo de pesquisas por mais de três décadas (VLAHOGIANNI, KARLAFTIS e GOLIAS., 2014), e está relacionada a uma área fundamental de ITS, conhecida como Sistema de Informação Avançado de Tráfego (ou ATIS, referente à *Advanced Traffic Information System*) (SUSSMAN, 2005) (GEORGESCU, ZEITLER e STANDRIDGE, 2012).

Uma etapa inicial importante para quem decide trabalhar com previsão de destino e rotas está relacionada com a seleção de dados de deslocamentos. Na literatura, é possível encontrar trabalhos que utilizam tanto dados sintéticos, como dados reais de rotas de usuários (VLAHOGIANNI, KARLAFTIS e GOLIAS., 2014). Dados reais, quando coletados durante vários dias, indiscutivelmente, apresentam um cenário mais representativo dos passeios feitos em uma cidade ou região, o que pode não ocorrer com dados sintéticos.

Atualmente, capturar dados reais de posicionamento ficou mais viável graças à possibilidade de utilizar o dispositivo GPS embutido nos atuais aparelhos celulares, por exemplo *smartphones*, ao invés de ter que investir em um dispositivo GPS separadamente. Tecnicamente, capturar um dado real de posicionamento equivale à obtenção de uma coordenada geográfica (latitude, longitude e altitude) associada a uma informação temporal (data e hora).

Outro ramo de pesquisa que se beneficiou com a facilidade de aquisição de pontos geográficos foi o de serviços baseados em localização (LBS, referente à *location-based services*) (PERERA *et al*., 2013). Serviços baseados em localização têm a característica de adicionar informação conceitual (ou contextual) ao posicionamento geográfico de determinado objeto (que pode ser um *smartphone*, um veículo ou uma ave, por exemplo) (SILVA, MACÊDO e CASANOVA, 2014). Algumas informações contextuais podem ser dia da semana, horário, regiões geográficas de origem e destino, tipo do POI, clima, entre outros.

Problemática Abordada nesta Tese

De maneira objetiva, foram identificadas três lacunas importantes relacionadas à previsão de trajetórias, e que são abordadas pelos modelos de previsão propostos nesta tese. São elas:

- A primeira lacuna está relacionada com a própria proposição de modelos de previsão. Uma parte dos modelos de previsão propostos utilizam apenas informações de lugares para previsão, sem considerar o conteúdo do deslocamento (por exemplo, os segmentos que são percorridos). Prever, também, os segmentos que serão percorridos é útil, por exemplo, para alertar sobre rotas congestionadas e para replanejar a previsão, caso um desvio seja tomado;
- A segunda lacuna em aberto é a possibilidade de prever trajetórias com base apenas em trajetórias anteriores. Assim, quando o usuário percorre um trajeto novo, há dois problemas: (1) a não identificação de que o trajeto é novo; e (2) prever um lugar que o usuário certamente não alcançará, por ser totalmente novo;
- A terceira lacuna refere-se ao uso das informações contempladas para previsão de trajetórias. Grande parte dos modelos existentes utiliza apenas informações geográficas e temporais para previsão, desconsiderando uma informação semântica importante: o papel que o lugar representa para o usuário (se *casa* ou *trabalho*, por exemplo). Saber o papel que um lugar representa permite que seja previsto um lugar não visitado anteriormente, mas com base no perfil de visitas do usuário. Por exemplo, considere que, aos sábados, no horário entre 19h e 20h, um usuário costuma visitar sempre o mesmo restaurante (a *lazer*), e, em certo momento, este usuário resolva conhecer outro restaurante, neste mesmo dia da semana e horário. Caso o usuário altere a sua ida para outro restaurante nunca visitado, um modelo de previsão que contempla a informação de papel do lugar pode identificar que, naquele dia da semana e horário ele costuma ir a um restaurante, e pode prever um restaurante próximo à sua localização, ainda que seja a primeira visita.

Escopo e Contribuições Principais desta Tese

Embora o foco do trabalho proposto seja prever destino e rotas, informações contextuais (inclusive, o papel que um lugar representa ao usuário – como *casa* ou *trabalho*) da localização são usadas para aperfeiçoamento dos modelos de previsão aqui propostos. Tais informações são especialmente importantes para previsão de destinos onde os usuários não estiveram anteriormente. No entanto, o escopo deste trabalho não investiga técnicas inovadoras no tocante aos mecanismos de identificação de papel que um lugar representa para o usuário, podendo, assim, ser considerado como um componente à parte sem prejuízo do detalhamento do *modelo de previsão*.

Na pesquisa ora apresentada, foi desenvolvido um sistema de previsão de rotas e destinos denominado de *Predroute*, que contempla os seguintes procedimentos: (1) identificar regiões de paradas automaticamente, isto é, lugares que o usuário permanece estacionário; (2) obter rotas que conectam as origens e os destinos dos trajetos, também de forma automática; (3) enriquecer semanticamente um trajeto, por meio da identificação do papel que um lugar representa para o usuário (por exemplo, se *casa* ou *trabalho*); e (4) prever rotas e destinos.

Embora o *Predroute* possua procedimentos para identificar regiões de paradas e obtenção de rotas, o principal foco da pesquisa reside no procedimento de **previsão de rotas e destinos**. As principais contribuições oriundas da pesquisa realizada são:

1. Desenvolvimento de um modelo de previsão com base na técnica de compressão de dados *Prediction by Partial Matching* (PPM), que possibilite o armazenamento de todos os segmentos que compõem a rota percorrida pelo usuário, e que realize a previsão de forma automática e ubíqua (isto é, sem a necessidade de uma interação ativa do usuário);
2. Desenvolvimento preliminar de um modelo de previsão que combina a técnica PPM, proposta no item (1), com as *Cadeias de Markov*. Assim, a previsão é realizada não somente com base no PPM, mas também pela influência das probabilidades de transição de estados das cadeias de Markov geradas;
3. Desenvolvimento preliminar de um modelo de previsão que combina a técnica PPM com as *Cadeias Ocultas de Markov* (HMM, do inglês *Hidden Markov Model*). Assim, a previsão é realizada combinando os cálculos realizados pelo PPM juntamente com o HMM;

4. Desenvolvimento de um mecanismo capaz de identificar o desvio de uma previsão realizada inicialmente. Ou seja, à medida que o usuário se desloca, o mecanismo detecta se o usuário está indo para o destino inicialmente previsto ou se aquela rota o levará a um destino nunca antes realizado. Denominamos este mecanismo de *rota reduzida*.

Os modelos de previsão propostos são elaborados de forma personalizada e individual, e são capazes de prever, inclusive, **lugares não visitados** pelos usuários, isto é, lugares que não foram considerados durante a aprendizagem. Um **lugar não visitado** refere-se a um lugar que pode existir, mas que o usuário não visitou, portanto, não está contemplado no conjunto de lugares visitados por ele. A verificação de existência deste lugar não visitado por um usuário pode ser feita mediante consulta a um conjunto de lugares externos ao modelo de previsão, como, por exemplo, mediante o uso do conjunto de lugares disponíveis no *Google Maps* ou no *Foursquare*. Assim, caso o usuário esteja seguindo a um lugar de destino que não conste no conjunto de lugares de seu histórico, será feita uma consulta ao conjunto de lugares externos ao modelo de previsão. Caso este lugar também não seja encontrado no conjunto de lugares externos, os modelos de previsão propostos não serão capazes de prever corretamente o destino a ser alcançado pelo usuário.

1.1. Motivação e Justificativa

Conforme discutido por WINTER et al. (2011), ITS possui uma vasta área de pesquisa a ser explorada, com muitos problemas em aberto, em que é possível contemplar pesquisas que abordam hardware e/ou software. Dentre estas, estão problemas referentes ao desenvolvimento de novos dispositivos para comunicação e integração de veículos; à obtenção de conhecimento adequado a partir da coleta de grande quantidade de dados de variadas fontes de informação; ao desenvolvimento de sistemas que envolvam questões referentes ao tráfego; entre outras. Portanto, oportunidades de pesquisas oriundas da ITS, relativas à área de computação, têm recebido atenção de pesquisadores nos últimos anos.

Com relação às oportunidades de pesquisa em ITS, de especial interesse neste trabalho estão as questões ligadas à previsão de rotas e destino. Atenção especial é direcionada às abordagens de previsão realizadas de forma automática, isto é, sem participação ativa do usuário no processo, como, por exemplo, para informar se este se

encontra em descanso ou trabalhando. Esta característica tem se mostrado pertinente com relação à usabilidade diária, uma vez que permite maior utilidade e transparência da tecnologia ao usuário. Acredita-se que um sistema que requer interação constante do usuário pode ser útil nos primeiros percursos, mas deixa de ser usado rotineiramente pela pessoa.

Prever o destino que um usuário alcançará e a rota que será percorrida, o mais cedo possível após o início de um deslocamento, é importante em variadas ocasiões. De posse desta informação, um sistema computacional pode identificar em quais vias há maior congestionamento e sugerir que o usuário as evite. Pode, também, sugerir que o usuário realize um desvio, em virtude de insegurança em algum trecho da rota, ou para visitar algum POI (como a ida a uma padaria ou farmácia), antes de alcançar o destino final.

Na tarefa de previsão de destino, é necessário ter um conjunto de deslocamentos realizados previamente para realizar previsões confiáveis. Em um deslocamento em curso de um usuário, prever um destino que ele já visitou antes é mais fácil do que prever um lugar que ele nunca tenha ido. Isto ocorre pelo fato de que um preditor se molda ao histórico de percursos do usuário, e fica limitado em virtude daquele conjunto de destinos já visitados, geralmente, desconsiderando outros tipos de lugares nunca antes visitados. A maioria dos trabalhos encontrados na literatura aborda a previsão de destino apenas a locais que os usuários já tenham visitado, desconsiderando a possibilidade de previsão de um lugar que ele nunca foi.

A maior motivação para o desenvolvimento desta pesquisa está em realizar a previsão de um destino que um usuário nunca tenha ido. O alcance deste objetivo implica em um sistema ainda mais útil, visto que o usuário poderá usufruir dos benefícios de uma previsão de destino (desvio de rotas congestionadas ou inseguras), mesmo para um lugar onde ele nunca esteve.

Quanto ao uso da técnica de compressão de dados PPM, esta já foi utilizada com sucesso no contexto de classificação, como, por exemplo, para classificação de obras literárias (PAVELEC, OLIVEIRA, *et al.*, 2009), classificação de arritmia cardíaca (ANDREZZA, BORGES e BATISTA, 2015) e classificação de textura (HONORIO, BATISTA e DUARTE, 2009). Assim, mediante estes casos de sucesso, foi pensado em um mecanismo capaz de transformar uma trajetória em símbolos que compõem uma palavra, isto é, uma trajetória passa a ser representada por um conjunto de símbolos (ou

palavras). Em seguida, a técnica PPM é utilizada no processo de compressão destas palavras, resultando nas árvores de símbolos PPM, que representam a aprendizagem do perfil de deslocamentos do usuário. Com isso, os modelos de previsão desenvolvidos neste trabalho, com base no PPM, têm a possibilidade de serem estendidos para outros domínios de problemas, desde que seja possível representar o objeto que está sendo manipulado (nesta tese, uma trajetória) como um vetor de símbolos. O funcionamento do PPM é descrito na Seção 4.2.

1.2. Questões de Pesquisa

Para guiar o desenvolvimento da pesquisa oriunda deste trabalho e poder avaliar cientificamente os preditores propostos, algumas Questões de Pesquisa (QP) foram formuladas. Com o auxílio de ferramentas estatísticas, como a utilização de métricas de avaliação (por exemplo, *precisão, cobertura* e *medida-F*) e testes de hipótese (como a *Análise de Variância*, ou ANOVA), é possível medir o alcance das questões propostas. Esta subseção se destina a apresentar e explanar sobre as duas questões de pesquisa definidas na tese, a saber:

QP1 *– Existe diferença no resultado do uso dos modelos de previsão de rotas e destino (implementados neste trabalho), incluindo previsão de lugares nunca visitados, com relação às métricas estatísticas (precisão, cobertura e medida-F) utilizadas para avaliação?*

H1-0: Não há influência entre os resultados obtidos, com relação ao uso dos modelos de previsão implementados, referentes às métricas estatísticas utilizadas para avaliação.

H1-1: Há influência entre os resultados obtidos, com relação ao uso dos modelos de previsão implementados, referentes às métricas estatísticas utilizadas para avaliação

Caso a hipótese nula seja refutada, significa que há diferença entre os resultados dos modelos de previsão propostos, e avaliados, nesta pesquisa. Portanto, usar um certo modelo de previsão, em detrimento de outro, resultará em variações nas variáveis dependentes analisadas, que são as métricas de precisão, cobertura e medida-F.

QP2 *– Existe diferença no resultado da previsão de trajetórias em uma base de dados com rotas que foram realizadas mais frequentemente (isto é, onde o par <origem, destino> foi realizado pelo menos duas vezes) versus uma base de dados que possui mais*

rotas que foram realizadas uma única vez (isto é, onde o par <origem, destino> foi realizado apenas uma vez) para os modelos de previsão baseados em PPM?

H2-0: Não há diferença no resultado da previsão de trajetórias em usar bases com rotas frequentes *versus* bases com rotas que contenham muitos trajetos realizados apenas uma vez para previsão de trajetória, referentes às métricas estatísticas de *precisão, cobertura* e *medida-F*.

H2-1: Há diferença no resultado da previsão de trajetórias em usar bases com rotas frequentes *versus* bases com rotas que contenham muitos trajetos realizados apenas uma vez para previsão de trajetória, referentes às métricas estatísticas de *precisão, cobertura* e *medida-F*.

Caso a hipótese nula seja refutada, significa que os estilos de deslocamentos (se rotas mais frequentes ou realizadas apenas uma vez, na sua maioria) de uma base de dados influenciam na previsão de destino, com relação às métricas estatísticas precisão, cobertura e medida-F. Assim, confirmada a refutação da hipótese nula, será possível segmentar uma previsão, por exemplo, em rotas esporádicas e rotas mais rotineiras, desenvolvendo preditores apropriados para cada um destes segmentos.

É importante mencionar, já neste momento, que cada questão de pesquisa foi analisada sob cenários diferentes. Por exemplo, para a **QP1**, a partir da mesma base de deslocamentos utilizada para testes, foram criados dois cenários de teste: (1) uma para as rotas que foram realizadas uma ou mais vezes, denominada de *RR1+*, que consiste no cenário que utiliza todos os deslocamentos da base de dados, sem realização de qualquer filtragem (o que dificulta a tarefa de previsão); e (2) outra para as rotas que realizadas duas ou mais vezes, denominada de *RR2+*, que consiste na seleção da base de dados das rotas realizadas duas ou mais vezes. Já para a **QP2**, os testes foram agrupados em três subseções, sendo uma para o modelo de previsão com base no PPM, outra para o modelo de previsão PPM-Markov e outra para o modelo de previsão PPM-HMM.

1.3. Objetivos

Nesta seção, é apresentado o objetivo geral do trabalho, além dos objetivos específicos.

1.3.1. Objetivo Geral

Desenvolver modelos de previsão de destino e rotas inovadores, que sejam capazes, inclusive, de prever lugares nunca visitados pelo usuário.

1.3.2. Objetivos Específicos

- Investigar e desenvolver modelos próprios e inovadores baseados em PPM para melhoria da previsão de destino e rotas;
- Implementar, de forma preliminar, outros modelos de previsão de rotas e destinos (nesta pesquisa, os que são baseados em cadeias de Markov e HMM) já consolidados pela literatura na área do conhecimento desta pesquisa, para comparação com os modelos propostos;
- Desenvolver um sistema que dê subsídios para o funcionamento adequado dos modelos de previsão, com implementação de componentes para *identificação de regiões de paradas*, *obtenção de rotas* a partir de dados espaço-temporais e *descoberta do papel que um lugar* representa para o usuário;
- Selecionar uma base de deslocamentos pública e disponível para acesso, para avaliação dos modelos de previsão propostos;
- Realizar avaliação experimental do modelo com dados reais, medir a eficácia do modelo com métricas estatísticas coerentes, como *precisão*, *cobertura* e *medida-F*, e compará-lo com os resultados disponíveis na literatura.

1.4. Publicações

À medida que a pesquisa deste doutorado foi avançando, a elaboração e a submissão de artigos foram sendo realizadas, para obter contribuições advindas dos revisores e também para verificar o posicionamento do modelo de previsão na literatura. No Quadro 1, são apresentadas as publicações realizadas até o momento, e qual melhoria foi acrescentada que justificou a publicação de um novo artigo.

1.5. Organização do Documento

Este capítulo contemplou as principais contribuições obtidas neste trabalho, além de motivar a necessidade desta pesquisa e em quais contextos estão sua utilidade. No

Capítulo 2, são apresentados e delineados os conceitos básicos adotados pelo trabalho. No Capítulo 3, uma revisão da literatura com os procedimentos e as formas de previsão abordadas pelos trabalhos correlatos são fornecidas. Os modelos de previsão propostos nesta tese de doutorado estão detalhados no Capítulo 4. No Capítulo 5, são apresentados os experimentos conduzidos nesta pesquisa, nos quais são descritos e analisados os resultados oriundos dos experimentos, referentes às questões de pesquisa formuladas neste Capítulo. No Capítulo 6, é apresentada a conclusão do trabalho desta tese e são apresentadas propostas para pesquisas futuras.

Quadro 1 - Catálogo de publicações.

Título do artigo	Ano	Veículo	Resumo/Melhoria
Predicting Routes and Destinations of Urban Trips using PPM Method	2015	Anais do 7o Simpósio Brasileiro de Computação Ubíqua e Pervasiva (SBCUP)	- Utilização do PPM para previsão de rotas e destino; - Uso de base de dados própria, com oito participantes.
Prediction of Destinations and Routes in Urban Trips with Automated Identification of Place Types and Stay Points	2015	Anais do XVI Brazilian Symposium on Geoinformatics Também publicado na Revista Brasileira de Cartografia (RBC), 2016	- Adição de uso de semântica pelo modelo; - Previsão também do tipo do lugar; - Uso de base de dados própria, com oito participantes.
A user-personalized model for real time destination and route prediction	2016	Anais do IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)	- Utilização do PPM para previsão de rotas e destino, incluindo lugares nunca visitados; - Apresentação do mecanismo de rota reduzida; - Utilização de base de dados pública para testes; - Uso de métricas estatísticas de *precisão*, *cobertura* e *medida-F*; - Criação de dois cenários para teste.

Fonte: Elaborada pelo autor.

Capítulo 2

Conceitos Básicos

Neste capítulo, são apresentados os conceitos básicos utilizados nesta pesquisa de doutorado. Definições importantes utilizadas ao longo de todo o documento são detalhadas, como é o caso da caracterização de: coordenadas geo-temporais; deslocamentos geográficos; segmentos, rotas, rotas parciais e rotas restantes; trajetórias anotadas semanticamente e de informações contextuais. Será esclarecido, também, que, no contexto deste trabalho, embora deslocamento, rota e trajetória tenham significados conceituais semelhantes (isto é, representam uma movimentação de um objeto entre uma origem e um destino), tecnicamente estes termos são semanticamente diferentes.

Ainda neste capítulo, serão descritos o funcionamento da técnica de compressão de dados *Prediction by Partial Matching* (PPM), das *Cadeias de Markov* e das *Cadeias Ocultas de Markov* (HMM), que são utilizados para o desenvolvimento de modelos para a previsão de rotas e destino. Adicionalmente, há um tópico que trata sobre agrupamento espacial, mecanismo utilizado neste trabalho.

2.1. Trajetórias

As principais questões relacionadas à trajetória são explanadas nesta seção, tais como: métodos para coleta de coordenadas geográficas (ou geo-temporais); definições sobre trajetórias geográficas e trajetórias que incorporam semântica; a importância de se identificar regiões de parada (regiões onde um usuário visita). É apresentada, ainda, uma técnica para reduzir o quantitativo de dados geográficos a ser manuseado por um modelo, conhecida por *map-matching* (QUDDUS e NOLAND, 2006).

2.1.1. Métodos para Coleta de Dados

A base de deslocamentos utilizada neste trabalho para avaliação dos modelos de previsão foi construída com o auxílio de dispositivos GPS (BRUSH, KRUMM e SCOTT, 2010), em que foram coletadas coordenadas geo-temporais de 21 usuários durante três meses, na cidade de Seattle – Estados Unidos.

Definição 1 Coordenada Geo-temporal - *Uma coordenada (ou ponto) geo-temporal **P** representa uma coordenada geográfica (latitude, longitude) com a informação temporal (data e hora) de sua captura.*

Embora dispositivos GPS tenham sido utilizados como método de coleta para a criação da base de deslocamentos utilizada nesta pesquisa, há outros métodos possíveis para obtenção de coordenadas geo-temporais. ZHANG *et al.* (2011) sumarizam os principais meios de coleta e suas respectivas aplicações, a saber:

- Dispositivos *Global Positioning System* (GPS): técnica de coleta adequada quando se quer saber a localização precisa onde o usuário está, principalmente quando o usuário está em um lugar aberto;
- Radares à laser: técnica adequada para auxílio de condutores a estacionamento de veículos; e detecção de pedestre;
- Sensores sísmicos: técnica mais indicada para classificação do tipo de veículo;
- Redes móveis de celulares: técnica de coleta adequada quando a precisão do local de um usuário não é tão importante, mas apenas a região geográfica onde ele está;
- Sensores meteorológicos: técnica útil para identificar a qualidade do ar, pressão atmosférica, temperatura e umidade;
- Redes Wi-Fi: técnica de coleta importante para captura de localização do usuário em lugares fechados.

2.1.2. Deslocamento Geográfico

Conceitualmente, as terminologias: deslocamento geográfico, rota e trajetória podem se referir a uma movimentação de objeto entre dois ou mais lugares (representados, por exemplo, por coordenadas geo-temporais). No entanto, no âmbito deste trabalho, estas terminologias possuem especificidades, sendo necessárias definições formais de cada um destes termos.

Definição 2 Deslocamento geográfico - *Um deslocamento geográfico D é representado por um conjunto de coordenadas geo-temporais ordenadas temporalmente (P_1, P_2, ..., P_n), em que P_n foi capturado após P_{n-1}, P_1 representa a origem do deslocamento e P_n representa o destino*.

Um deslocamento geográfico, no contexto deste trabalho, representa a movimentação de um usuário entre dois lugares, em que um conjunto de coordenadas geo-temporais, ordenadas de forma temporal, conecta estes dois lugares, denominados origem e destino. A *origem* de um deslocamento é uma região onde um objeto permanece parado até que uma ação de movimento tenha sido realizada. O *destino* de um deslocamento é a região onde um objeto irá ficar parado, após ter sido identificada alguma ação de movimento, estando precedida por uma *origem*. Lidar com um conjunto de coordenadas tem uma desvantagem principal: o grande quantitativo de dados a ser manipulado. Esta desvantagem ocorre, principalmente, quando a taxa de captura do posicionamento é elevada (alguns segundos), e muitas coordenadas são capturadas próximas umas das outras. Além disso, quando o usuário permanece parado durante muito tempo em um lugar, o posicionamento capturado pode ser praticamente igual. Para lidar com a problemática da elevada quantidade de dados GPS que pode ser coletada, o modelo desenvolvido nesta pesquisa converte um deslocamento geográfico representado por coordenadas geo-temporais para um conjunto de *segmentos*.

Definição 3 Segmento - *Um segmento S compreende exatamente duas coordenadas geo-temporais ($P_{início}$, P_{fim}), em que $P_{início}$ representa a coordenada de origem do segmento e P_{fim} representa a coordenada final do segmento (SIMMONS et al., 2006*).

Um único segmento consegue representar um número elevado de coordenadas geo-temporais, o que diminui consideravelmente o quantitativo de dados a ser manipulado por um modelo. Por exemplo, caso o usuário esteja em algum tráfego congestionado, várias coordenadas geo-temporais serão capturadas muito próximas. Ao invés de se manipular ou armazenar um grande número de coordenadas, pode-se, apenas, utilizar um segmento que represente diversas coordenadas. Assim, um modelo demandará menos memória computacional ao lidar com um conjunto de segmentos do que lidar com um conjunto de coordenadas geo-temporais.

Definição 4 Rota - *Uma rota R compreende uma sequência de segmentos (S_1, S_2, ..., S_n), em que $n > 0$ e S_n representa o n-ésimo segmento de uma rota.*

Tanto a terminologia *rota* como *deslocamento* representam, conceitualmente, a movimentação entre uma origem e um destino realizada por um usuário. No entanto, tecnicamente, uma *rota R* representa uma movimentação na forma de um conjunto de segmentos $(S_1, S_2, ..., S_n)$, enquanto um *deslocamento D* representa uma movimentação na forma de um conjunto de coordenadas geo-temporais $(P_1, P_2, ..., P_n)$, em que $|D| \geq |R|$ (a cardinalidade do conjunto de coordenadas geo-temporais é maior ou igual à cardinalidade do conjunto de segmentos).

2.1.3. Redução de Dados de Movimentação com Map-Matching

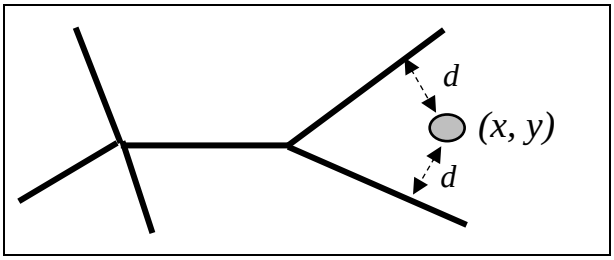
O procedimento para transformar um conjunto de coordenadas geo-temporais em uma sequência de segmentos é conhecido como *map-matching*, cuja definição é apresentada a seguir.

Definição 5 Map-matching – *É uma associação de uma coordenada geo-temporal a um segmento*, e *pode ser caracterizada como uma função* $\boldsymbol{f}$ *que associa uma coordenada geo-temporal* $\boldsymbol{P}$ *a um segmento* $\boldsymbol{S}$. *Assim, temos* $\boldsymbol{f: P \rightarrow S}$.

O *map-matching* é uma técnica que associa coordenadas geo-temporais a segmentos, com a finalidade de reduzir o quantitativo de dados de movimentações a ser manipulado por um modelo (QUDDUS e NOLAND, 2006). O processo de associação que ocorre no *map-matching* lida com situações consideravelmente complexas. Por exemplo, na Figura 1, existe uma coordenada geográfica entre dois segmentos (que poderiam representar duas ruas, por exemplo), e pode ser difícil identificar o segmento correto ao qual a coordenada deve estar associada.

É possível resolver este problema verificando a orientação do deslocamento do usuário e os segmentos recentemente percorridos, analisando a qual dos dois segmentos conflitantes uma coordenada deve ser associada. É possível, também, corrigir uma associação incorreta de uma coordenada a um segmento, verificando os segmentos posteriores. Isto é, se o provável segmento seguinte (S_{n+1}) não for continuação imediata do segmento corrente (S_n), provavelmente, S_n foi escolhido erroneamente, e precisa ser corrigido.

Figura 1 - Problema de associação da coordenada geográfica (x,y) ao segmento correto.



Fonte: Elaborada pelo autor.

2.1.4. Identificação de Regiões de Paradas

Ao ordenar temporalmente as coordenadas geo-temporais capturadas através de dispositivos GPS, é possível identificar dois padrões principais nestes dados: (1) um padrão referente ao deslocamento entre duas regiões geográficas; e (2) um padrão que representa as regiões geográficas visitadas. O padrão de dados do grupo (1) representa um *deslocamento geográfico* (**Definição 2**), que deverá ser convertido para uma *rota* (**Definição 4**). Já o padrão de dados do grupo (2) representa uma *região de parada*, que são as regiões (ou lugares) visitadas pelo usuário.

Definição 6 Região de parada - *Uma região de parada (RP), ou apenas parada, é uma região geográfica onde um usuário permaneceu estacionário por uma duração maior que um limiar de tempo T (em minutos). Isto é, RP representa a tupla <P, D, Δt >, em que P é uma coordenada geográfica (latitude e longitude), D delimita a região em torno do ponto P (raio), e Δt é o tempo que o usuário permaneceu na região ($\Delta t > T$). O conceito de permanecer parado significa que as coordenadas capturadas durante um período maior que T não são distantes entre si por uma distância maior que um limiar D (em metros).*

Nesta pesquisa, o conceito "*permanecer estacionário*" significa que uma pessoa permaneceu em uma região de parada. Portanto, um quantitativo alto de captura de posicionamento em uma mesma região geográfica é convertido para uma *região de parada*, caso os requisitos de distância e duração sejam atendidos. Na Seção 4.2.1, são apresentados o procedimento adotado nesta tese para identificação de regiões de paradas e a escolha dos valores adotados para os limiares de tempo *T* e distância *D*. Há dois motivos para que regiões de paradas possam ser identificadas, a saber: (1) identificar, a partir de um conjunto de coordenadas geo-temporais, o que representa deslocamento e o que representa uma região de parada; e (2) adicionar semântica às movimentações, uma

vez que, ao se identificarem as regiões de paradas, pode-se sugerir o papel que o local representa a uma pessoa (como, por exemplo, *trabalho* ou *lazer*).

2.1.5. Enriquecimento Contextual às Movimentações

Utilizar informação semântica pode influenciar positivamente modelos que manipulam trajetórias de diversas maneiras (NANNI *et al.*, 2010). No contexto de previsão, saber a semântica a respeito dos lugares visitados por um usuário pode ser útil, por exemplo, para prever um lugar para onde um usuário nunca tenha ido. Neste trabalho, adicionar semântica a uma trajetória significa associar um *papel do lugar* à origem e um *papel do lugar* ao destino, e usa-se o termo *trajetória anotada semanticamente* para se referir às trajetórias em que a origem e o destino estão associados aos *papéis do lugar*.

Definição 7 Trajetórias anotadas semanticamente - *Uma trajetória anotada semanticamente, ou apenas trajetória semântica, é uma trajetória à qual foram adicionadas informações mais detalhadas sobre sua realização. Tipos de informações semânticas incluem o meio de transporte utilizado; o* ***papel do lugar*** *(por exemplo, casa, trabalho, lazer); e o tipo do POI (por exemplo, restaurante, centro comercial) (PARENT et al, 2013).*

Neste trabalho, o meio de transporte considerado pelos participantes é o seu próprio veículo. O tipo do POI poderá ser usado para investigações futuras, mas não é considerado neste momento. A variável dinâmica atualmente considerada por esta pesquisa é o *papel do lugar*, ou seja, o papel que um lugar tem para uma pessoa. Assim, uma mesma região geográfica (um *centro comercial*, por exemplo) pode representar um lugar para *lazer* para um usuário, e para *trabalho* a outro.

Além de semântica, outro recurso que pode aperfeiçoar um modelo de modo que este possa prever mais apropriadamente uma rota e/ou destino é o uso de informação contextual. Contexto diz respeito aos fatores (ou variáveis) ambientais que determinado objeto sob análise está sujeito à interferência (VIEIRA, TEDESCO e SALGADO, 2011). De maneira geral, ao se analisar contexto, é possível obter informações relevantes acerca de um ambiente que está em torno de um objeto principal. A partir de informações relevantes coletadas, é possível identificar comportamentos diferentes de um mesmo objeto ou processo, dependendo das condições (ambientes) onde o objeto se encontra. As variáveis contextuais a serem consideradas em um estudo são definidas conforme o objeto

em análise. Como o foco deste trabalho é trajetória, mais especificamente previsão, a *informação contextual* acerca de um trajeto considerada é apresentada na **Definição 8**.

Definição 8 Informação contextual - *Informação contextual representa os dados sobre uma movimentação geográfica realizada. Neste trabalho, dia da semana e intervalo de tempo da partida, origem e destino, além do papel do lugar, são variáveis consideradas contextuais.*

As informações contextuais referentes às rotas realizadas por um usuário são importantes para delinear o perfil da pessoa além de sua movimentação geográfica. Neste trabalho, são consideradas as seguintes variáveis contextuais:

- **Dia da semana** de uma partida da origem, representada por um conjunto de números inteiros finitos, em que os inteiros são associados aos seguintes valores: 0 = domingo; 1 = segunda-feira; 2 = terça-feira; 3 = quarta-feira, 4 = quinta-feira; 5 = sexta-feira; 6 = sábado.
- **Intervalo de tempo, ou hora,** de uma partida da origem, em que tal intervalo é representado por um inteiro que corresponde a um intervalo *i* entre dois horários (*0, para* $0 < i \leq 1$*; 1, para* $1 < i \leq 2$*; ...;* 23*, para* $23 < i \leq 24$);
- (Região geográfica de) **Origem** e **destino**, que representam, respectivamente, os lugares de início e fim de uma rota.

2.1.6. Previsão de trajetórias

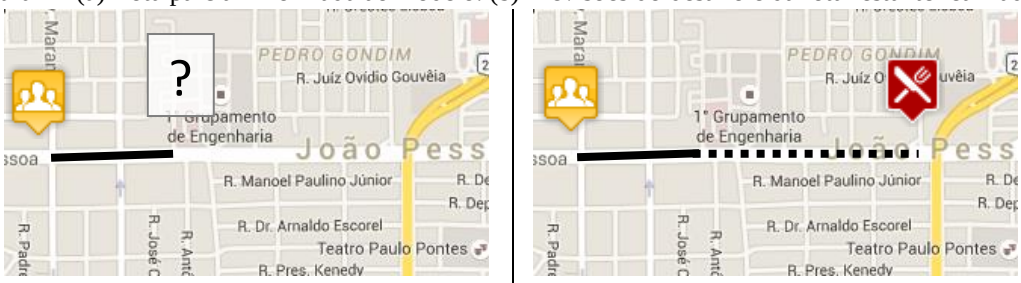
Para previsão de trajetórias, no âmbito desta pesquisa, existem os conceitos de rota parcial, rota restante e trajetória. Embora os conceitos de *trajetória* e *rota*, em algumas ocasiões, possam representar a mesma ação (movimentação de um lugar para outro), tecnicamente eles apresentam diferenças. Uma *trajetória*, conforme **Definição 11**, é composta por uma *rota* (ver **Definição 4**), por *informações contextuais* (ver **Definição 8**) e por *semântica* (ver **Definição 7**).

Definição 9 Rota parcial - *Uma rota parcial P representa um subconjunto de segmentos* ($S_1, S_2, S_3, ..., S_m$) *de uma rota R* ($S_1, S_2, S_3, ..., S_m, ..., S_n$, *em que* $1 \leq m < n$), *em que* S_m *representa o m-ésimo segmento de uma rota em curso (isto é, o objeto está em movimento e ainda não alcançou seu destino).*

Definição 10 Rota restante - *Uma rota restante F (S_{m+1}, S_{m+2}, ..., S_p, S_n, onde $p > m$ e $p + 1 \leq n$) representa um subconjunto de segmentos previstos para o usuário chegar a um destino, em que S_m representa o m-ésimo segmento de uma rota em curso, S_n representa o segmento onde se encontra o destino da rota e S_p representa o segmento anterior ao segmento S_n e posterior ao segmento S_p.*

Como entrada para o modelo de previsão, é informada uma rota geográfica parcial, juntamente com o identificador do usuário e as informações contextuais da movimentação. A saída do modelo é uma trajetória prevista, com informações sobre o destino e o papel do lugar que o usuário está indo. A Figura 2 (a) ilustra a rota parcial dada como entrada, isto é, o deslocamento que já foi percorrido, representado pela linha sólida. A Figura 2 (b) apresenta o resultado da previsão para a trajetória informada, representada pela linha pontilhada.

Figura 2 - (a) Rota parcial informada ao modelo. (b) Previsões do destino e da rota restante realizadas.



(a) (b)

Fonte: Elaborada pelo autor.

Definição 11 (Trajetória) *Uma trajetória é um objeto que compreende uma rota realizada juntamente com suas informações contextuais. Formalmente, uma trajetória é uma tupla $T = <s, i, o, d, l_o, l_d, u, R>$, em que:*

- ***s** representa o dia da semana da trajetória;*
- ***i** representa o intervalo de tempo do início da trajetória;*
- ***o** representa a região geográfica do início da trajetória;*
- ***d** representa a região geográfica do destino da trajetória;*
- ***l_o** representa o papel do lugar do início da trajetória;*
- ***l_d** representa o papel do lugar do destino da trajetória;*
- ***u** representa o identificador do usuário; e*
- ***R** representa a rota (geográfica).*

Os modelos de previsão propostos nesta pesquisa, conforme será detalhado no Capítulo 4, apresentam, como resultado do processamento, uma trajetória prevista. Prever trajetória, neste contexto, significa que serão previstos um destino e uma rota geográfica (até o destino), além das questões contextuais. Portanto, será previsto, também, o papel do lugar que um usuário deve ir, já que uma trajetória engloba informações espaço-temporais, além de contextuais.

2.2. Prediction by Partial Matching (PPM)

A técnica de compressão de dados *Prediction by Partial Matching* (PPM), baseada em modelos estatísticos, é uma das mais sofisticadas no quesito compressão, e também uma das mais eficientes em compressão de dados sem perda de informação (SALOMON, 2004). A ideia principal deste método é usar uma árvore de símbolos adaptativa em um contexto finito, em que a compressão de um símbolo é dada pela probabilidade de sua aparição nos últimos *n* símbolos, ao invés da probabilidade de ocorrência em toda a fonte de informação.

Quando um símbolo está na iminência de ser codificado, é verificada sua probabilidade de ocorrência no contexto de maior ordem, isto é, nos últimos *n* caracteres. Caso seja encontrado, o símbolo será codificado com a probabilidade corrente e, após a codificação, a probabilidade será atualizada. Caso não seja encontrado, um símbolo especial, conhecido como *escape*, é codificado e tentar-se-á codificar o símbolo no contexto (de menor ordem) dos últimos *n-1* símbolos. Este procedimento é repetido até que o símbolo a ser codificado seja encontrado em alguma das ordens (últimos *n-i* símbolos, em que $0 \leq i \leq n$). O contexto com menor ordem é o de tamanho 0 (zero), que contém todos os símbolos já codificados da fonte de informação. A probabilidade de ocorrência do símbolo em um contexto é igual à frequência de sua aparição na fonte, dividida pela soma do quantitativo de símbolos diferentes mais o quantitativo de símbolos já codificados até o momento naquele contexto. O quantitativo de símbolos diferentes ocorrido em um contexto representa a frequência do símbolo especial *escape*. A variante do PPM usada neste trabalho foi o PPM-C (MOFFAT, 1990), já que o *escape* é o recurso utilizado para chavear de um contexto de maior ordem para um de menor ordem.

No contexto deste trabalho, uma árvore PPM é utilizada para representar uma rota. Para isso, cada segmento é identificado de forma única, por meio de um valor inteiro. A

lista de segmentos que representa uma rota é convertida para um vetor de inteiros, que equivale a uma cadeia de símbolos a ser codificada pelo PPM. Quando houver essa codificação, o algoritmo gerará uma árvore de símbolos, conhecida como árvore de símbolos PPM ou árvore PPM (ver Tabela 1).

A Figura 3 apresenta uma rota particionada em uma lista de segmentos, em que cada um deles está rotulado com um número inteiro. No rodapé da Figura 3, há um vetor de números inteiros que será informado como entrada para a criação de uma árvore PPM (ver Tabela 1). A figura ilustra um hipotético deslocamento que o usuário realizou de *casa* para o *trabalho*. Considere, também, que esse deslocamento *casa* → *trabalho* foi realizado nove vezes, e a árvore PPM gerada está disponível na Tabela 1. Contextualizando o funcionamento do PPM no exemplo de uma trajetória, e conforme a Figura 3, toda vez que o usuário iniciar um deslocamento de casa passando pelos segmentos 'D' e 'F', o próximo segmento provável é o 'G' (com uma probabilidade de 0.9). No entanto, se o usuário não seguir pelo segmento 'G', o caracter de *escape* (ou simplesmente *esc*) é ativado para chavear o contexto de menor ordem (contexto 'F'), com probabilidade 0.1.

No âmbito desta pesquisa, a etapa de treinamento consiste em obter as árvores de símbolos PPM para cada rota diferente realizada pelo usuário. Já a etapa de teste consiste em codificar uma dada rota a ser prevista com as árvores PPM já armazenadas, criadas na etapa de treinamento. Quando uma rota é informada para teste (isto é, um vetor de inteiros), esta é codificada com as árvores PPM, e o algoritmo gerará como resposta uma medida chamada de *Razão de Compressão* (RC). O valor de RC é um indicador de similaridade: quanto maior seu valor, mais similar a rota a ser testada é com uma árvore PPM. Uma característica do PPM que é útil para modelos preditivos é a capacidade de elaborar rapidamente uma árvore de símbolos, modelada conforme a fonte de informação.

Figura 3 - Deslocamento: inicia de casa, atravessa oito segmentos e tem como destino o trabalho.

Fonte: Elaborada pelo autor.

2.3. Modelos de Markov

Esta subseção explana sobre os dois modelos de Markov utilizados neste trabalho: (1) as cadeias simples de Markov; e (2) os modelos Ocultos de Markov (Hidden Markov Models, ou HMM). Ambos os modelos foram implementados para avaliação experimental e comparação preliminares com os preditores propostos neste trabalho, a saber: o modelo baseado em PPM; o modelo PPM-Markov, que combina o uso de cadeias de Markov com PPM; e o modelo PPM-HMM, que combina PPM com HMM.

Tabela 1 - Árvore de símbolos que representa o deslocamento da Figura 3.

Ordem *k = 2*				Ordem *k = 1*				Ordem *k = 0*				Ordem *k = -1*			
Predições		c	p	Predições		c	p	Predições		c	p	Predições		c	p
→	F	9	$^9/_{10}$	A →	D	9	$^9/_{10}$	→	A	9	$^9/_{90}$	→	A	1	$^1/_{\|A\|}$
→	*Esc*	1	$^1/_{10}$	→	*Esc*	1	$^1/_{10}$	→	D	9	$^9/_{90}$				
DF →	G	9	$^9/_{10}$	D →	F	9	$^9/_{10}$	→	F	9	$^9/_{90}$				
→	*Esc*	1	$^1/_{10}$	→	*Esc*	1	$^1/_{10}$	→	G	9	$^9/_{90}$				
								→	H	9	$^9/_{90}$				
FG →	H	9	$^9/_{10}$	F →	G	9	$^9/_{10}$	→	I	9	$^9/_{90}$				
→	*Esc*	1	$^1/_{10}$	→	*Esc*	1	$^1/_{10}$	→	J	9	$^9/_{90}$				
GH →	I	9	$^9/_{10}$					→	P	9	$^9/_{90}$				
→	*Esc*	1	$^1/_{10}$	▪▪▪				→	*Esc*	9	$^9/_{90}$				

Fonte: Elaborada pelo autor.

2.3.1. Cadeias de Markov

O modelo de Markov é categorizado como parte de modelos estocásticos, isto é, possui a capacidade de modelar transições que se comportam de maneira randomizadas em um sistema com estados observáveis. A forma mais simples de um modelo de Markov é uma cadeia de Markov, em que a probabilidade de transição entre os estados depende apenas do estado atual de um objeto ou processo. Assim, um modelo de Markov possui a característica de ser um *processo sem memória* (ou *memoryless process*), já que o passado de estados percorrido é desprezado. Outra característica é que nas cadeias de Markov o espaço de estados é discreto, ao invés de ser contínuo (NORRIS, 1998). Assim, a

transição de um estado *A* no tempo *t-1* para outro estado *B* no tempo *t*, depende apenas da probabilidade condicional de *B* em relação a *A*, ou *P(B | A)*. De modo mais formal, a **probabilidade de transição** entre os estados é dada da seguinte forma,

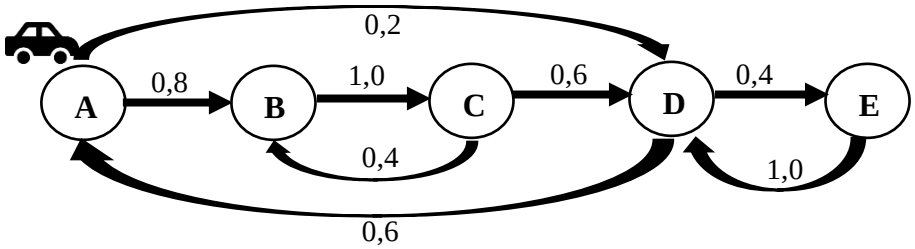
$$Pi,j = P(X_{t+1} = j \mid X_t = i)$$

em que:

- *i* e *j* representam dois estados quaisquer do espaço de estados *E*;
- X_t representa o estado que o objeto ocupa no instante de tempo *t*; e
- $P_{i,j}$ representa a **probabilidade de transição** do estado *i* no instante *t* para o estado *j* no instante de tempo *t+1*.

A Figura 4 apresenta um objeto (por exemplo, um carro) que pode realizar transições entre cinco estados possíveis {A, B, C, D, E}. O objeto está inicialmente no estado A e tem probabilidade de 0,8 para ir ao estado B, e 0,2 para ir ao estado D. Do estado B, o objeto sempre se desloca para o estado C (probabilidade de transição do estado B para o C é 1). As demais distribuições de probabilidade de transição entre os estados da Figura 4 estão disponíveis na Tabela 2. As linhas da tabela representam o estado atual onde o objeto se encontra, as colunas representam os estados possíveis de serem alcançados e as células armazenam os valores de probabilidade que o objeto tem de transicionar de estado na linha *i* para um estado na coluna *j*. O somatório das probabilidades para cada uma das linhas resulta em 1.

Figura 4 - Representação e probabilidade da transição entre os espaços de estados.



Fonte: Elaborada pelo autor.

Tabela 2 - Distribuição de probabilidades do espaço de estados da Figura 4.

		Estado no instante de tempo *t+1*					
		A	*B*	*C*	*D*	*E*	*Soma*
Estado no instante de tempo *t*	*A*	0	0,8	0	0,2	0	1
	B	0	0	1	0	0	1
	C	0	0,4	0	0,6	0	1
	D	0,6	0	0	0	0,4	1
	E	0	0	0	1	0	1

Fonte: Elaborada pelo autor.

2.3.2. Modelo Oculto de Markov

Uma cadeia de Markov é representada pela tupla $<E, Q, M>$, em que E é o espaço finito de estados, Q representa a quantidade finita de ações e M representa a matriz de probabilidade de transição $E \times E$. Enquanto uma cadeia simples de Markov considera as probabilidades de transição de um objeto entre seus espaços de estados apenas baseado no estado corrente (ou nos últimos n estados), um Modelo Oculto de Markov (ou *HMM*, referente a *Hidden Markov Model*) considera aspectos referentes ao estado inicial do objeto no instante de tempo t e às *sequências de observações* ocorridas para descobrir o *estado oculto* que um objeto deverá estar nos próximos instantes de tempo ($t+1$, $t+2$, ..., $t+k$) (STAMP, 2015).

Uma tupla em um Modelo Oculto de Markov é representada por $<E, A, O, M, B, \pi>$, em que E, A e M representam as mesmas informações de uma cadeia de Markov, O representa um conjunto finito de observações, B representa a matriz de probabilidade de observações $E \times O$ e π é a distribuição de probabilidade inicial dos estados. Assim, se $E = \{e1, e2\}$ e $O = \{o1, o2, o3\}$, tem-se que $|A| = 4$ (em virtude de $E \times E$), $|B| = 6$ (em virtude $E \times O$) e $|\pi| = 2$, referente aos dois estados de E.

A probabilidade de transição entre os estados da matriz A pode ser dada na forma $a_{i,j}$, em que i representa o estado corrente e j o estado de destino, isto é, $a_{i,j} = P(\text{estado } e_j \text{ em } t+1 \mid \text{estado } e_i \text{ em } t)$. A probabilidade de observações representada pela matriz B pode ser dada da seguinte forma $b_j(k)$, em que j é o estado corrente e k é a observação (por exemplo, uma das observações do conjunto $\{o1, o2, o3\}$). Assim, formalmente, tem-se que $b_j(k) = P(\text{observação } k \text{ em } t \mid \text{estado } e_j \text{ em } t)$.

Para melhor entendimento, considere um exemplo genérico com as características a seguir: um espaço de estados $E = \{e_1, e_2\}$; o seguinte conjunto de possibilidades de observações $O = \{O_1, O_2\}$; e uma matriz π de probabilidade inicial. Dada uma sequência de estados aleatória $X = (e_1, e_2, e_1, e_1)$ e uma sequência aleatória de observações $Y = (o1, o2, o1, o2)$, a probabilidade $P(X, Y)$ de obtenção da sequência X dada a sequência de observações Y é dada da seguinte forma,

$$P(X, Y) = \pi_{e1} b_{e1}(O_1) a_{e1,e2} b_{e2}(O_2)\ a_{e2,e1} b_{e2}(O_1)\ a_{e1,e1} b_{e2}(O_2) \qquad \text{Equação (2)}$$

Considere um contexto específico de aplicação de um HMM no deslocamento do usuário entre três lugares que ele costuma visitar (estados) dependendo das condições climáticas (observação). O espaço de estados $E = \{casa, academia, praia\}$ representa os

lugares, enquanto o conjunto de observações *O = {sol, chuva}* representa as condições climáticas. A matriz de probabilidades de transição de estados, representada por *A*, é

$$A = \left(\begin{array}{l|lll} & casa & academia & praia \\ \hline casa & 0,1 & 0,6 & 0,3 \\ academia & 0,6 & 0 & 0,4 \\ praia & 0,8 & 0,2 & 0 \end{array} \right),$$

a matriz de observação, representada por *B*, é

$$B = \left(\begin{array}{l|ll} & sol & chuva \\ \hline casa & 0,4 & 0,6 \\ academia & 0,2 & 0,8 \\ praia & 0,9 & 0,1 \end{array} \right),$$

e a distribuição de probabilidade inicial, representada por π, é

$$\pi = \left| \begin{array}{lll} 0,5 & 0,3 & 0,2 \end{array} \right|.$$

Ao utilizar HMM para descobrir uma sequência de observações *Y = {sol, chuva, sol}*, deve-se utilizar a Equação (2), e calcular a probabilidade da sequência de observações *Y* para cada combinação de estados de *E*. Por exemplo, a probabilidade para a obtenção dos estados *casa-academia-casa* é representada por *P(casa, academia, casa)*, e seu cálculo é obtido da seguinte maneira:

P(casa, academia, casa) = 0,5(0,4)(0,6)(0,8)(0,6)(0,2) = 0,01152

Após calcular a probabilidade para as demais combinações, o HMM deverá encontrar as probabilidades normalizadas para os estados na primeira posição de observação, depois para a segunda posição de observação e, por fim, para a terceira. Isto é, deve-se calcular a probabilidade de *casa*, *academia* e *praia* na primeira posição de observação, depois as posições dois e três das observações. O HMM escolhe o estado que tiver a maior probabilidade na primeira posição, depois o estado que obtiver a maior probabilidade na segunda posição e assim por diante. O resultado do processamento do

HMM será a descoberta (previsão) dos *modelos ocultos*, obtidos indiretamente pelo cálculo da probabilidade de transição entre os estados juntamente com as observações.

É importante mencionar os três problemas fundamentais que podem ser resolvidos utilizando um HMM (JURAFSKY e H. MARTIN, 2008), a saber:

- **Problema 1**: Dado um HMM $\lambda = (A, B, \pi)$ e uma sequência de observações *Y*, procura-se determinar a probabilidade do aparecimento da sequência *Y* dado o modelo λ, isto é, encontrar $P(Y \mid \lambda)$;
- **Problema 2**: Dado um HMM $\lambda = (A, B, \pi)$ e uma sequência de observações *Y*, procura-se determinar melhor sequência de estados conforme a sequência *Y* dado o modelo λ, isto é, encontrar $P(Y, \lambda)$;
- **Problema 3**: Dada uma sequência de observações *Y*, o espaço de estados *E* e o conjunto de observações possíveis *O*, encontrar o modelo $\lambda = (A, B, \pi)$ que maximize a probabilidade das observações *Y*.

2.3.2.1. Modelo Oculto de Markov no Contexto desta Pesquisa

No contexto desta pesquisa de doutorado, as *observações O* de uma Cadeia Oculta de Markov correspondem aos lugares visitados por um usuário, às informações contextuais de dia da semana (sete valores para os sete dias da semana) e ao intervalo de tempo (24 valores para os 24 intervalos de tempo referentes a um espaço de hora). O espaço de *estados E* representa os lugares que o usuário visitou. Portanto, uma vez que as colunas de uma matriz representam as *observações* de um HMM, o quantitativo de colunas é obtido pela multiplicação da quantidade de lugares visitados (*L*), que é variável dependendo do usuário, pela quantidade de dias da semana (*sete dias*) pela quantidade de horários (*24 horas*), resultando no cálculo $L \times 7 \times 24$. Já as linhas da matriz representam o espaço de estados, ou seja, os lugares que o usuário visitou.

2.4. Agrupamento Espacial

Um dos procedimentos usados para identificação e mineração de características de dados espaciais é o *agrupamento espacial*. Uma técnica de agrupamento espacial é capaz de reunir (agrupar) objetos conforme alguma classe de característica, como a proximidade espacial e/ou temporal, e pode ser do tipo *baseada em distância* (*distance-*

based), hierárquica, baseada em densidade (density-based) e *baseada em grade* (HAN, KAMBER e TUNG, 2001).

Uma técnica *baseada em distância* recebe como entrada um parâmetro principal: o quantitativo de grupos a ser criado. A necessidade desta variável referente ao quantitativo apresenta duas desvantagens: (1) geralmente, o número de grupos a ser criado depende dos dados de entrada, e não há como prever este valor até os dados serem processados; e (2) todos os pontos geográficos **devem** ser alocados a um dos grupos, o que pode não ser útil quando se quer identificar um deslocamento geográfico (KISILEVICH *et al*., 2010). Um exemplo de técnica *baseada em distância* é o algoritmo K-Means (KANUNGO *et al*., 2002). Uma técnica *baseada em densidade* precisa das informações de distância mínima entre os pontos (um raio) para identificar um grupo e do quantitativo mínimo de pontos para uma região ser considerada como um agrupamento (TORK, 2012). Uma vantagem desta técnica comparada àquela está no fato de que a última identifica tanto os pontos de um grupo quanto os pontos que são considerados *ruídos*. Um ruído, na verdade, representa pontos espaciais que não fazem parte de um agrupamento, ou seja, é um ponto que é candidato a fazer parte de uma movimentação geográfica.

Como exemplos de algoritmos da técnica *baseada em densidade*, há aqueles que consideram somente a distância espacial, como o DBSCAN (ESTER *et al*., 1996). No DBSCAN, é considerada apenas informação geográfica, isto é, coordenadas geográficas que residam próximas umas das outras são candidatas a criação de agrupamentos, independentemente de quando foram capturadas. Isto pode não ser útil no contexto de trajetórias, já que, se um usuário realiza a mesma trajetória todos os dias, pode ser que uma avenida seja erroneamente considerada como um grupo pelo DBSCAN, por ter sido desconsiderada a informação temporal. Para contemplar informações temporais, foi proposto o *Spatial-Temporal* DBSCAN (BIRANT e KUT, 2007) (ou ST-DBSCAN). Neste algoritmo, a criação de um grupo de pontos espaciais deve obedecer às restrições de espaço (estarem distantes até uma distância *D*) e de tempo (estarem próximos até um tempo *T*).

No modelo proposto nesta pesquisa, é utilizado um algoritmo para identificação de grupos (ou *pontos de paradas*) *baseado em densidade*, mais especificamente no algoritmo ST-DBSCAN. Isto é, até a fase de processamento dos dados de deslocamento, não é possível identificar quantos grupos de pontos de paradas serão criados. Este número

só será definido após o procedimento de *Identificação de Paradas* do modelo proposto (ver Seção 4.2.1). Os dados de deslocamento devem estar ordenados temporalmente antes que o procedimento de identificação de pontos de paradas seja iniciado.

Quanto ao *método hierárquico* de agrupamento espacial, este cria uma hierarquia entre os grupos encontrados. A descoberta dos grupos espaciais pode ser realizada de duas maneiras: (1) *aglomerativa* (abordagem *bottom-up*), em que os objetos são alocados a grupos previamente definidos, e, a partir destes grupos, vão sendo criadas hierarquias; ou (2) *divisiva* (abordagem *top-down*), em que todos os objetos são agrupados em um único grupo e, a cada iteração, o grupo de maior nível vai sendo dividido em grupos de menores níveis com os objetos sendo alocados a estes grupos mais específicos. O método de agrupamento espacial baseado em *grade* define previamente o quantitativo de grupos em forma de grades, isto é, quantiza o espaço em um número finito (e pré-definido) de células que forma uma estrutura de grade. Esta técnica surgiu da necessidade de aperfeiçoar a performance do método baseado em *densidade*. No entanto, há perda de informação, uma vez que, dependendo da dimensão da grade, mais de um grupo diferente de objetos pode fazer parte de uma mesma grade (HAN, KAMBER e TUNG, 2001).

2.5. Conclusão do Capítulo

Este capítulo apresentou os conceitos básicos relacionados a trajetórias a serem considerados no restante do trabalho. Foram esclarecidas as diferenças entre os conceitos de *deslocamento*, *rota* e *trajetória*, importantes para o entendimento do restante do documento. Foi apresentada, ainda, a técnica de compressão de dados PPM, componente principal do modelo para a previsão de trajetórias, bem como foi feita uma breve descrição sobre conceitos acerca de agrupamento espacial. Além do mais, foram feitas explanações sobre as Cadeias de Markov e o Modelo Oculto de Markov, que foram combinados com PPM, resultando em modelos de previsão de trajetórias inovadores. No capítulo seguinte, serão apresentados os trabalhos relacionados a esta pesquisa.

Capítulo 3

Revisão da Literatura

Neste capítulo, são apresentadas as pesquisas mais importantes e com contribuições pertinentes para a área de previsão de trajetória. A maioria dos trabalhos foca no problema de previsão de trajetória sem considerar aspectos semânticos (por exemplo, papel do lugar da origem e do destino de um deslocamento), desconsiderando uma informação que pode ser útil para o desenvolvimento de um modelo preditor mais robusto. Embora pesquisas na linha de previsão não sejam recentes (NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM, 2012), integrar conhecimento semântico às trajetórias tem recebido atenção apenas nos últimos anos (SPACCAPIETRA et al., 2008). Devido a isto, este capítulo dedica uma seção para tratar da linha de investigação sobre previsão sem usar informação semântica e outra seção que apresenta trabalhos que contemplam informação semântica para a melhoria da previsão.

3.1. Previsão de Trajetórias sem Semântica

Para previsão de trajetórias, o modelo criado por Simmons et al. (2006) usa a técnica de *Cadeias de Markov Oculta* (HMM) e informação contextual (dia da semana, horário e velocidade do veículo) em um corpus formado por 46 rotas na região de Michigan, nos Estados Unidos. A taxa de acerto do modelo deles é de 98%, embora seja necessário fazer uma ponderação: apenas 5% das transições entre os segmentos permite seguir por mais de um caminho. Isto é, 95% dos segmentos tem apenas um único caminho à frente, o que pode facilitar a tarefa de previsão do próximo segmento. Para o caso dos 5% das transições que permitem seguir por mais de um segmento, a taxa de acerto do

algoritmo fica entre 70% e 80%. O modelo preditor desenvolvido por Krumm (2008) prevê apenas os próximos segmentos (até os 10 seguintes), sem prever o destino de um trajeto. Seu modelo utiliza *Cadeias de Markov* para previsão, analisa os 10 últimos segmentos percorridos e, com uma taxa de acerto de 90%, consegue prever o próximo segmento. Quando é feita a previsão para os próximos 10 segmentos, a taxa de acurácia diminui para 50%. O modelo proposto por este trabalho, prevê, além do próximo segmento, também o destino assim que uma viagem inicia.

O modelo proposto por Froehlich e Krumm (2008) usa a técnica de *Algoritmo de Casamento de Padrão* (do inglês *Closest Match Algorithm*), baseado no algoritmo de *distância de Hausdorff* (HUTTENLOCHER e RUCKLIDGE, 1992), que tenta identificar a similaridade entre uma rota em curso (rota parcial) e uma rota realizada anteriormente. Se houver uma rota similar (interseção e proximidade entre as coordenadas geográficas do deslocamento), recupera-se a rota restante, baseada em uma rota já realizada. O modelo proposto por eles alcança uma precisão de 40%, quando 50% da rota é percorrida. Uma limitação para a aplicabilidade prática do trabalho deles, porém, é a ausência da técnica de *map-matching* (já explicada na Seção 2.1.3). Por outro lado, Tiwiri et al. (2013) apresentam uma metodologia similar àquela de Froehlich e Krumm (2008), no entanto, utilizando a técnica de *map-matching* para reduzir o tamanho dos dados. No trabalho publicado por eles, a taxa de acurácia foi semelhante à de Froehlich e Krumm (2008), porém, o quantitativo de dados manuseado foi reduzido e a eficiência do modelo foi melhor.

O algoritmo PPM foi proposto para o problema de previsão de lugares no trabalho de Burbey e Martin (2008). O modelo de previsão proposto pelos autores considera informações contextuais referentes à chegada do usuário aos lugares que costuma visitar, a duração que o usuário permanece nestes lugares e o lugar atual que ele se encontra. Com base nestas informações, é previsto o próximo destino do usuário, mas sem considerar os segmentos a serem percorridos. Uma vez que o preditor desenvolvido por Burbey e Martin (2008) considera apenas a origem do usuário, não há mudança no destino a ser previsto, mesmo quando o usuário segue por outro caminho.

Xue et al. (2013) desenvolveram um modelo que usa *cadeias de Markov* e *inferência Bayesiana* para previsão de rotas. A abordagem permite prever, inclusive, lugares que um usuário não visitou ainda, ou seja, superando a limitação de prever passeios apenas baseado em dados históricos. Isto é possível pelo fato de que uma rota

realizada gera uma ou mais rotas novas (não realizadas), para contemplar regiões circunvizinhas à rota que foi percorrida. O corpus utilizado no trabalho deles para testes foi de motoristas de taxi, cujos deslocamentos são definidos pelos passageiros. Embora o modelo de previsão proposto por Xue et al. (2013) consiga prever lugares nunca visitados, ele faz uso apenas de informações geográficas, e não considera o papel que um lugar representa para o usuário (como, por exemplo, *casa* ou *trabalho*).

Lee et al. (2016) desenvolveram um preditor que utiliza *Casamento de Padrões* para prever o próximo destino, dada uma origem. Uma trajetória é representada por uma tupla que contém informações sobre a localização da origem, localização do destino, o dia da semana e o tempo que o usuário permaneceu lá. Esta trajetória é chamada de *Spatiotemporal-periodic* (STP). Após isso, o preditor segmenta uma trajetória em subtrajetórias, para poder identificar trechos esporádicos e rotineiros. Assim, uma trajetória será composta por um conjunto de subtrajetórias, em pode possuir lacunas. Esse conjunto de subtrajetórias é chamado de GSTP, no trabalho deles. Assim, mesmo que o usuário percorra uma rota esporádica, o algoritmo de *Casamento de Padrões*, por meio das trajetórias GSTP, associa o trajeto corrente com algum no passado.

Herder et al. (2014) apresentaram várias abordagens de previsão de destino, realizando uma comparação entre elas. Foram implementados preditores baseados nas técnicas:

- *Top-N*: em que o preditor tenta prever o destino baseado nos *N* lugares mais visitados;
- *Last-N*: em que o modelo tenta prever o destino baseado nos últimos *N* lugares que o usuário visitou;
- *Hora*: em que o modelo seleciona os *N* lugares de um mesmo horário que a previsão de destino está sendo feita, e tenta prever conforme esse horário;
- *Distância*: em que o modelo seleciona os *N* lugares mais próximos da posição corrente do usuário para prever o destino;
- *Cadeia Simples de Markov*: este modelo utiliza uma cadeia simples de Markov, em que observa o estado (lugar) atual para prever o destino.

Para avaliação das abordagens implementadas, foram utilizadas as bases de deslocamentos *MSR GPS Privacy* (BRUSH, KRUMM e SCOTT, 2010), mesma base de dados que é utilizada para avaliar os algoritmos de previsão propostos nesta tese; e a base

do projeto do *GeoLife* (ZHENG *et al*., 2009). Conforme os experimentos apresentados pelos autores, as cadeias de Markov obtiveram os melhores resultados, alcançando uma precisão de 62.9%.

Trasarti et al. (2015) propuseram um preditor que mescla as informações individuais e coletivas dos usuários para realizar previsão de destino. Assim, uma vez que são usadas informações acerca de deslocamentos de todos os usuários para predição, o modelo deles é capaz de prever lugares ainda não visitados. É importante destacar que o modelo deles manipula coordenadas geo-temporais, ao invés de realizar qualquer discretização (por exemplo, uso de *map-matching*), o que pode gerar uma grande quantidade de dados para manuseio. Os autores alegam que o uso de coordenadas geo-temporais resultará em uma previsão de destino (para eles, o destino é uma coordenada geo-temporal) espacialmente mais próxima do real. Para predizer uma rota, eles utilizam um algoritmo que representa uma função de distância da rota que está em curso com as rotas históricas dos usuários. A rota histórica que obtiver o menor valor resultante do cálculo desta função, será considerada para a previsão.

Liu et al. (2016) desenvolveram um sistema individual de recomendação, que realiza a previsão dos *N* (*top-N*) destinos mais prováveis para onde o usuário irá se deslocar. A metodologia do preditor se baseia no método *Skip-gram*, em que ocorre a associação de uma palavra diferente para cada lugar diferente visitado pelo usuário, resultando em uma *sentença*. A *sentença* representa todos os lugares visitados em um trajeto, e ela é informada na etapa de treinamento do preditor. Para a sugestão dos destinos mais prováveis a serem visitados pelo usuário, é utilizada uma extensão da técnica *Weighted Approximately Ranked Pairwise* (WARP).

Técnicas de descoberta de conhecimento (*KDD*), como *regras de associação*, já foram usadas para criação de modelos preditores. Quando um veículo inicia um deslocamento, uma regra de associação é gerada para a rota em curso (conforme os segmentos percorridos). Em seguida, uma função de identificação de padrões é ativada, para verificar se o conjunto de segmentos percorrido até o momento está em uma árvore de segmentos gerada anteriormente. O modelo proposto por Morzy (2006) utiliza uma versão do algoritmo *Apriori* para geração das regras de associação. Tanaka et al. (2009) apresentam um procedimento de previsão baseado na via que o usuário está (como, por exemplo, se rodovia ou via arterial). Além disso, o processo considera também as informações contextuais de dia da semana, horário do deslocamento, número de

passageiros, condições climáticas e peso da bagagem. Em Monreale et al. (2009), uma trajetória é representada por um conjunto de células, em que cada célula representa as localizações visitadas pelo usuário. Associada a cada localização, está o horário que o usuário a visitou. Para previsão de destino, o modelo utiliza regras de associação.

Figueiredo et al. (2016) propuseram um modelo denominado de *TriberFlow*, que utiliza uma Cadeia *Semi Markoviana*, juntamente com a técnica *Passeio Aleatório* (do inglês, *Random Walk*). Na etapa de criação do modelo preditor, os trajetos são categorizados em dois grupos: (1) aqueles realizados frequentemente (também chamados de *trajetos estacionários*), e que, segundo os autores, são mais fáceis de prever; e (2) aqueles realizados esporadicamente (também chamados de *trajetos transientes*), isto é, trajetos que não são realizados rotineiramente. Com relação à informação contextual, o modelo faz uso da informação de data e hora de realização dos trajetos.

Para previsão de destino, também já foi usada a técnica de Árvores de Sufixo Probabilística, ou somente PST, do inglês *Probabilistic Sufix Trees* (RON, SINGER e TISHBY, 1996), que cria uma árvore de contexto para representar uma cadeia de Markov de ordem variável, semelhante ao que ocorre na técnica de compressão de dados PPM. No âmbito do trabalho de Lei, Li e Peng (2013), uma trajetória é representada por um conjunto de células, onde cada célula pode representar vários segmentos. O modelo de previsão de Lei, Li e Peng (2013) funciona com informações individuais e considera, como entrada, a informação geográfica referente ao deslocamento, além do horário do deslocamento. De posse destas informações, o modelo de previsão constrói as árvores de sufixo. Rocha et al. (2016) desenvolveram um *sistema* para previsão de destinos chamado *TPRED*, que também utiliza um mecanismo de previsão de destino baseado no PST. De maneira semelhante ao que ocorre no trabalho de Lei, Li e Peng (2013), uma trajetória também é representada por um conjunto de células. O sistema descrito pelos autores extrai e utiliza, além da informação geográfica, as informações contextuais: dia da semana e hora. O modelo de previsão proposto Rocha et al. (2016), além de predizer para onde o objeto em movimento se desloca (isto é, destino do objeto), também prediz quando o objeto deve alcançar o destino.

3.2. Uso de Semântica para Previsão de Trajetórias

Em sistemas baseados em localização, a informação semântica representa a associação de dados espaciais sobre a localização a uma informação que agrega valor característico sobre o local (SPACCAPIETRA *et al*., 2008). Isto é, busca-se associar dados espaciais, referentes aos pontos de paradas, ao papel que determinada parada possui para um usuário (se *casa* ou *trabalho*, por exemplo). Com isso, a semântica busca superar a limitação de identificar apenas os pontos de paradas, mas entender a razão pela qual o usuário visitou uma certa região geográfica, e realizou determinado deslocamento (se *casa* → *trabalho*).

O trabalho de Ying et al. (2011) está entre os pioneiros na utilização de semântica para melhoria da previsão de destinos. O corpus de deslocamento utilizado para testes foi criado por captura de dispositivos GPS e por sinais das torres de celulares. Foi criada uma base de anotações semânticas, conhecida como *Geographic Semantic Information Database (GSID)*, que contém informações oriundas do *Google Maps*[1]. O modelo de previsão apresentado se divide em dois módulos principais: um *offline*, responsável por anotar semanticamente as regiões geográficas de paradas; e outro *online*, que prevê em tempo real o destino que um usuário deve alcançar. A abordagem obteve uma taxa de acurácia entre 53% e 68% na previsão do destino. Ying et al. (2014) apresentaram um aperfeiçoamento, em que o novo modelo é capaz de sugerir itens a serem comprados, caso o usuário esteja em um estabelecimento.

Lung et al. (2014) desenvolveram um modelo para previsão de destinos e descoberta do meio de transporte utilizado. O modelo deles usa HMM para a tarefa de previsão e consulta à biblioteca do *Google Maps* para enriquecer semanticamente uma trajetória. Utilizando dados reais, o modelo de previsão obteve uma taxa de acurácia de 68.3% na previsão do próximo lugar a ser visitado. Uma lacuna identificada no trabalho publicado por Lung et al. (2014), no entanto, é a dificuldade de identificar a base de dados que foi utilizada nos testes (por exemplo, não está clara a duração da coleta, nem o meio de transporte utilizado). Além disso, não foi apresentado um detalhamento maior sobre o procedimento de como a etapa de previsão é realizada. Embora Lung et al. (2014) apresente uma abordagem que possibilite a previsão de lugares que o usuário nunca

[1] https://maps.google.com

visitou antes, este detalhamento fica vago. O que fica claro, apenas, é que eles usam semântica para tipos de lugares, e realizam detecção do meio de transporte utilizado.

Huang et al. (2012) desenvolveram um modelo capaz de observar quatro aspectos contextuais referentes à localização onde um usuário está, como informação temporal, espacial, comportamental e ambiental. O modelo deles utiliza variáveis contextuais e semânticas para realizar previsão do próximo tipo de lugar que um usuário vai estar, e foi elaborado para lidar com os dados disponíveis pelo desafio da Nokia, conhecido como *Mobile Data Challenge* (MDC) (LAURILA *et al*., 2012). O modelo de previsão desenvolvido por Huang et al. (2012) utilizou um extrator de características dos dados. Depois, foi utilizada a análise Chi-quadrado para identificar as características que possuem maior influência na taxa de acurácia (mais representativas). Para os testes, foram utilizados os algoritmos de classificação SMO (PLATT, 1998), J48 (QUILAN, 1993) e *SimpleLogistic* (PENG, LEE e INGERSOLL, 2002). Ocorreram testes tanto com os algoritmos de forma isolada, como também com a combinação entre eles. A melhor taxa de acurácia foi obtida combinando os algoritmos SMO e SimpleLogistic, em que foi obtido 65.77% de acurácia. Zhu et al. (2013) também desenvolveram um modelo para o desafio MDC, destinado para a inferência do tipo de lugar onde um usuário está. Para elaboração do modelo de inferência deles, foram usados os algoritmos de Lógica de Regressão (LR) (PENG, LEE e INGERSOLL, 2002), *Support Vector Machine* (SVM) (CORTES e VAPNIK, 1995), *Random Forests* (RF) (BREIMAN, 2001) e *Gradient Boost Trees* (GBT) (NATEKIN e KNOLL, 2013). A melhor taxa de acurácia obtida foi com o algoritmo GBT, com 75.1% de acerto.

O modelo de previsão desenvolvido por Nademgeba et al. (2012), conhecido como *Destination Prediction Model* (DPM), utiliza informações contextuais para aperfeiçoamento da previsão, incluindo o papel que um lugar representa para o usuário. As informações contextuais, no entanto, devem ser reportadas pelo usuário através de um questionário, tarefa esta que não é prática para uso rotineiro em modelos preditivos. Para a previsão, o modelo deles utiliza um algoritmo que representa uma função de distância, para obter a trajetória histórica mais próxima da trajetória atual. Além disto, dada a posição atual, o modelo realiza uma análise das informações contextuais do usuário (como horário e dia da semana) e observa os últimos lugares visitados pelo usuário (inclusive, o contexto destes lugares).

3.3. Análise dos Trabalhos Relacionados

Nos trabalhos descritos na Subseção 3.1, os modelos de previsão propostos realizam a previsão de destino e/ou rota considerando apenas informações espaciais e temporais, além de algum dado contextual, como dia da semana e horário da partida de um deslocamento. Outro aspecto a ser observado é que, na maior parte dos trabalhos, a trajetória é representada por um conjunto de células, ao invés de um conjunto de segmentos. Proceder desta maneira diminui o quantitativo de dados a ser manipulado por um modelo, porém, diminui a precisão espacial da previsão de rotas e de destino, especialmente quando o tamanho da célula é muito grande. Uma das diferenças do *Predroute* para os demais modelos de previsão apresentados na Subseção 3.1 está no uso de informações semânticas, isto é, o *Predroute* contempla a informação do papel que um lugar representa para o usuário.

Nos trabalhos apresentados na Subseção 3.2, os modelos de previsão consideram o papel que um lugar representa para o usuário apenas como uma informação complementar para a previsão de destino, mas não a utiliza na tentativa de predizer um destino nunca visitado. Uma diferença entre o *Predroute* e a maioria dos modelos de previsão descritos na Subseção 3.2 é a capacidade de identificar que um trajeto está sendo realizado pela primeira vez, por meio do mecanismo de *rota reduzida*, além de poder prever lugares que o usuário não tenha visitado.

As cadeias de Markov e HMM, que são métodos consolidados na literatura de previsão de destino, possuem limitações importantes, sendo, uma delas, a possibilidade de uso de elevado recurso computacional para processamento de informações (ABE e WARMUTH, 1992) (GILLMAN e SIPSER, 1994). Por exemplo, o processo de aprendizagem e inferência de uma cadeia de Markov de ordem variável, ou de um HMM, pode consumir muito recurso computacional, o que pode resultar na inviabilidade de se usar segmentos para representação de uma trajetória. Para lidar com esta problemática, a técnica PPM é apropriada, uma vez que requer menos recurso computacional, e representa uma cadeia de Markov de ordem variável, tornando-a possível de ser utilizada em um sistema disponível para uso comercial.

O Quadro 2 apresenta os principais trabalhos relacionados (descritos nas Subseções 3.1 e 3.2) com esta pesquisa, analisando alguns itens importantes que puderam ser selecionados durante a revisão da literatura. Na primeira coluna do Quadro, estão os

trabalhos sob análise. A partir da segunda coluna, estão as características consideradas para comparação destes trabalhos. As descrições de cada coluna estão dispostas logo abaixo do Quadro.

3.4. Conclusão do Capítulo

Este capítulo apresentou os principais trabalhos correlatos a esta tese de doutorado. Neste capítulo, os trabalhos foram agrupados em duas linhas de análise: uma que agrupa trabalhos que contemplam informações contextuais, mas sem considerar informação semântica; e outra que, além de contemplar informações contextuais, utiliza semântica dos lugares na previsão de trajetórias. Foi apresentada, ainda, uma análise crítica dos trabalhos relacionados e as lacunas de pesquisas em aberto nesta temática. No próximo capítulo, serão detalhados os modelos de previsão propostos nesta tese, além de apresentar a arquitetura do sistema *Predroute*.

Quadro 2 - Sumarização dos trabalhos relacionados.

Trabalho	Item A	Item B	Item C	Item D	Item E	Item F	Item G	Item H
Ying et al. (2011)	N	C	C	N	A	T	T	NA
Froehlich e Krumm (2008)	S	I	G	N	P	G	A	NA
Tiwiri et al. (2013)	S	I	S	N	P	G	A	NA
Burbey e Martin (2008)	N	I	SD	N	P	G	A	NA
Xue et al. (2013)	S	C	C	S	A	G	C/O	NA
Herder et al. (2014)	N	I	SD	N	A	G	P/O	NA
Simmons et al. (2006)	S	C	S	N	P	G	A	S
Krumm (2008)	S	I	S	N	P	G	A	NA
Huang et al. (2012)	NA	I	C	S	A	T	A	S
Zhu et al. (2012)	NA	NA	SD	S	A	T	A	S
Figueiredo et al. (2016)	NA	I	SD	N	A	G	A	NA
Nademgeba et al. (2012)	S	I	C	N	A	T	A	NA
Lee et al. (2016)	N	I	C	N	A	G	A	NA
Lie et al. (2016)	NA	I	SD	N	A	G	P/O	S
Trasarti et al. (2015)	S	C	G	S	P	G	P/O	NA
Lung et al. (2014)	S	I	NA	NA	NA	T	A	NA
Rocha et al. (2016)	NA	I	C	N	P	G	P/O	NA
Predroute (oriundo desta tese)	S	I	S	S	A	T	T	S

Fonte: Elaborada pelo autor.

Item A: Previsão em tempo real? Valores possíveis: Sim (S), Não (N) ou Não Informado (NA)
Item B: Modelo de previsão proposto é coletivo ou individual? Valores possíveis: Coletivo (C), Individual (I) ou Não Informado (NA)
Item C: Qual a granularidade da composição de uma rota? Valores possíveis: células/grade (C), segmentos (S), Coordenadas Geo-Temporais (G), Somente Destino (SD) ou Não Informado (NA)
Item D: Prevê lugares não visitados anteriormente? Valores possíveis: Sim (S), Não (N) ou Não Informado (NA)
Item E: Se dados de usuários reais, a base de rotas é própria ou pública? Valores possíveis: Própria (P) ou Pública/Aberta (A)
Item F: Utiliza que tipo de informação contextual? Geográfica, Temporal e Semântica (T); ou somente Geográfica e Temporal (G);
Item G: Quais são as métricas utilizadas para avaliação (precisão, cobertura e medida-F) do preditor? Valores possíveis: Todas as métricas, isto é, precisão, cobertura e medida-F, (T); Precisão (P); Cobertura (C); Medida-F (M); Acurácia (A); Outra (O); e Não Informado (NA);
Item H: Há validação cruzada? Sim (S), Não (N) ou Não Informado (NA)

Capítulo 4

Predroute: Um Sistema para Previsão de Destinos e Rotas

Este capítulo descreve a arquitetura do sistema *Predroute*, que contempla as implementações dos modelos de previsão propostos nesta tese, a saber: PPM, PPM-Markov e PPM-HMM. Além disso, neste capítulo, são apresentados, ainda, os principais componentes que compõem o *Predroute* e são explicadas as responsabilidades de cada um destes componentes. Na subseção referente à arquitetura do sistema, também há uma subdivisão em dois tópicos: (1) um que apresenta que explana sobre os principais componentes; e (2) outro que apresenta o esquema conceitual do banco de dados de trajetórias. Neste capítulo, há uma subseção responsável pelo detalhamento do componente de Previsão, que descreverá, de maneira conceitual, sobre os modelos de previsão de destinos e rotas propostos por esta pesquisa. Posteriormente, é apresentado, de forma algorítmica, o funcionamento dos modelos de previsão, através de pseudocódigo. Além disto, será explicada como ocorre a integração do PPM com as Cadeias de Markov (PPM-Markov) e com o HMM (PPM-HMM).

4.1. Arquitetura do Sistema Predroute

No diagrama de componentes do sistema *Predroute*, disponível na Figura 5, é possível identificar quatro componentes principais, a saber: (1) *previsão;* (2) *obtenção de rotas;* (3) *identificação de paradas;* e (4) *enriquecimento semântico*. Conforme ilustrado

na figura, os três componentes estão destacados na cor cinza não fazem parte da contribuição principal deste trabalho, porém, foram implementados para dar suporte ao componente de previsão. Para implementação destes três componentes, foram utilizados algoritmos preconizados pela literatura. O componente de *Previsão* recebe destaque, pois contempla as principais contribuições oriundas desta tese.

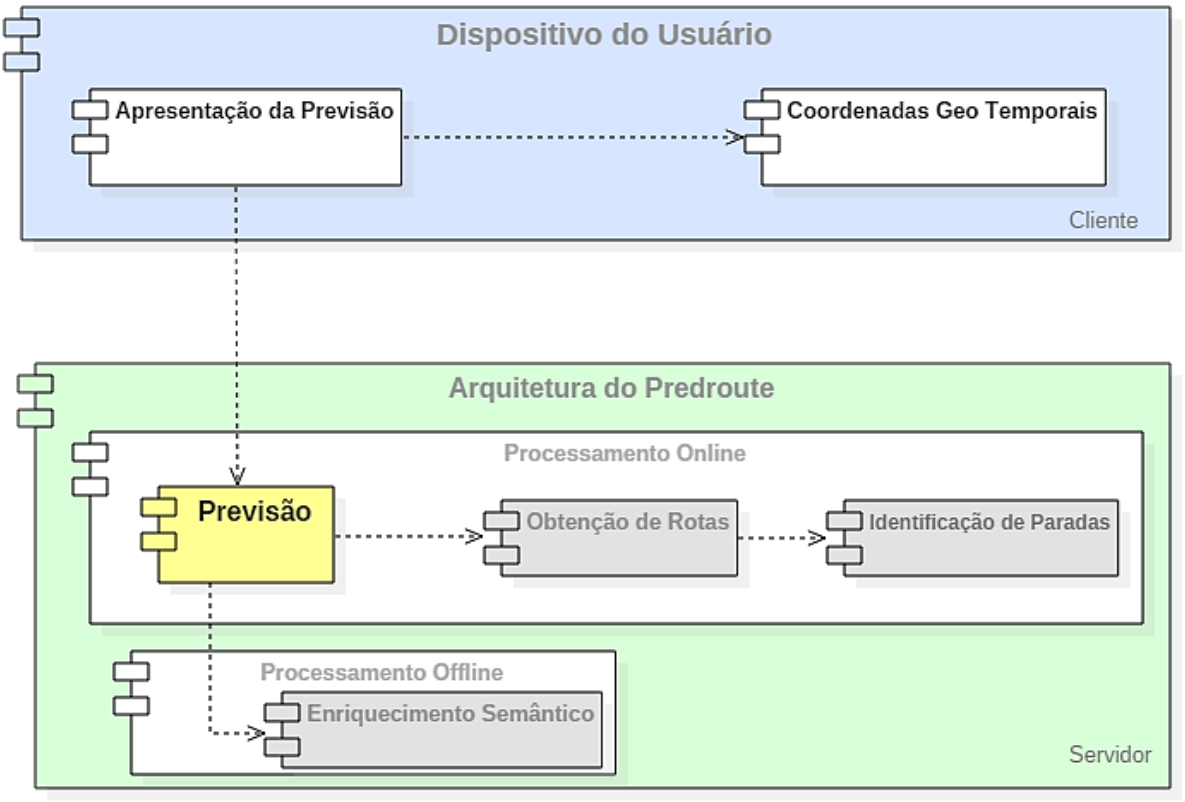
4.1.1. Componentes do Sistema Predroute

A Figura 5 ilustra o diagrama de componentes do *Predroute*, em que os componentes estão alocados conforme o local de funcionamento: se no dispositivo do usuário (cliente) ou se no servidor. O *Predroute* recebe como dados de entrada, para previsão, um conjunto de coordenadas geo-temporais juntamente com o identificador de um usuário. Como saída, o *Predroute* gera a previsão de uma trajetória, que contempla o destino, o papel do lugar daquele destino e a rota até chegar àquele local. Para que a previsão seja realizada no *Predroute*, faz-se necessário que os dados de deslocamentos, representados por coordenadas geo-temporais, sejam convertidos em um conjunto de segmentos. Para isso, foram criados os componentes para obtenção de rotas, identificação de regiões de paradas e enriquecimento semântico. Cada uma destas etapas possui um componente associado, que são definidas a seguir:

- **Previsão de trajetórias**, que tem a responsabilidade de prever o destino para onde o usuário deverá alcançar e a rota a ser percorrida. Este componente é composto pelas etapas de aprendizagem, em que são criados os padrões de deslocamentos individuais de trajetórias, e de previsão, que consiste na previsão de uma rota restante, dada uma rota parcial. Como entrada, este componente recebe um conjunto de coordenadas geo-temporais, e, como saída, gera a previsão de uma trajetória;
- **Obtenção de rotas (procedimento de *map-matching*)**, que tem a função de associar as coordenadas GPS de um deslocamento aos segmentos corretos. A utilidade deste módulo reside em diminuir o quantitativo de dados a ser manipulado pelo sistema *Predroute*;
- **Identificação de paradas**, que tem a responsabilidade de identificar, de forma automática, os lugares (ou regiões de parada) que os usuários visitam;

- **Enriquecimento semântico**, que tem a característica de identificar o papel que a origem e o destino representam para o usuário. Por exemplo, se a trajetória ocorreu de *casa* para o *trabalho* ou se de *casa* para o *lazer*. Este processo, assim como os anteriores, também é realizado sem a necessidade de interação do usuário com o modelo.

Figura 5 - Diagrama de componentes do sistema Predroute. A principal contribuição deste trabalho está no módulo **Previsão**. Os demais módulos, embora sejam importantes, não fazem parte do foco de pesquisa deste trabalho, e utilizaram algoritmos preconizados pela literatura.



Fonte: Elaborada pelo autor.

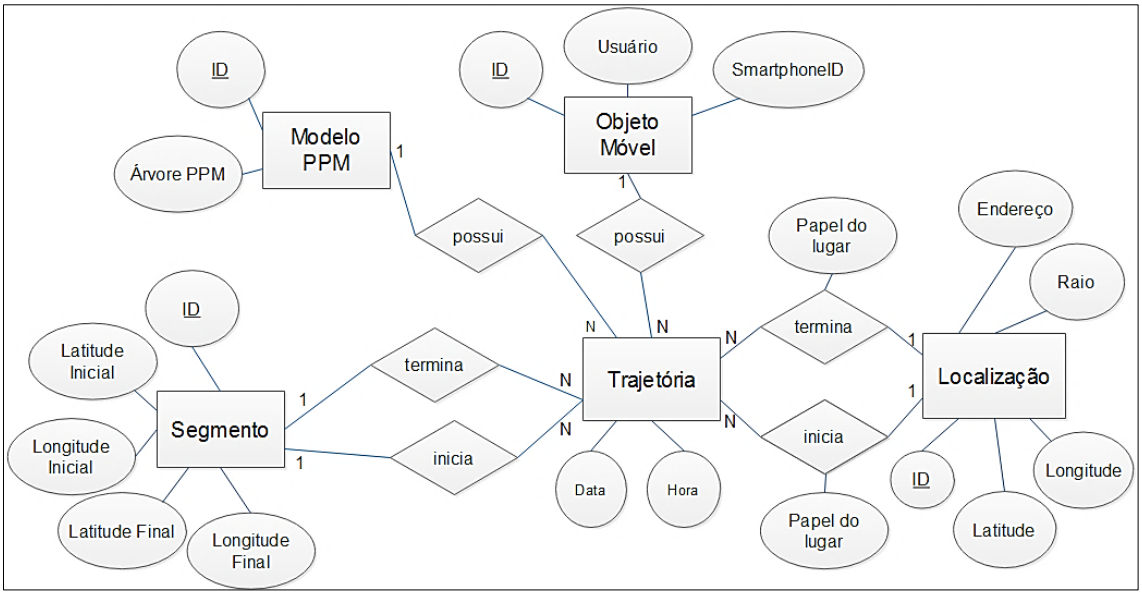
4.1.2. Modelo de Entidades e Relacionamentos

A modelagem conceitual do banco de dados (CHEN, 1976) do *Predroute*, que foi usada como referência para posterior definição dos projetos lógico e físico, está ilustrada na Figura 6. Percebe-se, pois, a existência de uma entidade *Trajetória*, responsável por representar uma trajetória realizada. A entidade *Trajetória* possui os atributos de *data* e *hora*, que representam a informação temporal sobre o trajeto, e possui relacionamentos com as entidades *Objeto Móvel*, *Modelo PPM*, *Segmento* e *Localização*.

A entidade *Objeto Móvel* representa o objeto que está contido em uma trajetória, sendo, no contexto deste trabalho, representado por um usuário. A entidade *Modelo PPM*

representa a árvore de símbolos PPM criada para toda trajetória diferente que um determinado usuário realiza. Toda rota realizada é composta por um conjunto de segmentos, em que o primeiro segmento representa o segmento de origem do trajeto e o último segmento representa o segmento de destino. Assim, a entidade *Segmento* representa os segmentos que o usuário percorreu na origem e no destino de uma rota. Em virtude disso, esta entidade possui dois relacionamentos com a entidade *Trajetória*, um que representa o início e outro que representa o final do trajeto. Cada segmento é composto pelas coordenadas espaciais (latitude e longitude) de início e fim. A entidade *Localização* representa as regiões de paradas (ou os lugares) visitadas pelo usuário. Como em toda trajetória há a região de parada inicial e a região de parada final, há dois relacionamentos entre a entidade *Trajetória* e a entidade *Localização*.

Figura 6 - Diagrama conceitual do banco de dados do Predroute.



Fonte: Elaborada pelo autor.

Na subseção seguinte, serão detalhados os modelos de previsão baseados em PPM, e serão apresentados os algoritmos implementados.

4.2. O Componente Previsão

Esta subseção inicia com a apresentação dos modelos de previsão baseados em PPM, propostos por esta tese. Após a apresentação dos modelos, serão descritos os algoritmos de previsão de rotas e destino.

4.2.1. Modelos de Previsão Baseados em PPM

Para lidar com os desafios da temática de previsão de rotas e de destino, foram elaborados três modelos de previsão de rotas e destino baseados na técnica de compressão de dados PPM. Com o uso desta técnica (PPM), é possível elaborar uma árvore de símbolos PPM que representa o conjunto de segmentos percorridos em uma rota. Cada árvore de símbolos criada representa uma, e somente uma, rota. Os três modelos de previsão propostos se baseiam em: (1) PPM combinado com Cadeia de Markov; (2) PPM combinado com HMM; e (3) PPM usado sem combinação com outra técnica.

Como parte destes modelos de previsão propostos, há um mecanismo para a conversão de um deslocamento geográfico em uma árvore de símbolos PPM. Para a criação desta árvore de símbolos, inicialmente, um conjunto de coordenadas geo-temporais, que representa um deslocamento, é convertido (com o auxílio de uma técnica de *map-matching*) para uma rota (um conjunto de segmentos). Este conjunto de segmentos é, em seguida, convertido em uma árvore PPM, obtido pelo processo de compressão do PPM. Embora, conceitualmente, o conjunto de segmentos e a árvore de símbolos PPM representem uma rota, o primeiro está na forma plana e o segundo está na forma comprimida, por ter sido submetido ao processo de compressão do PPM. Dividindo-se o tamanho do arquivo que está na forma plana pelo tamanho do arquivo que está na forma comprimida, obtém-se o valor referente à *Razão de Compressão* (RC).

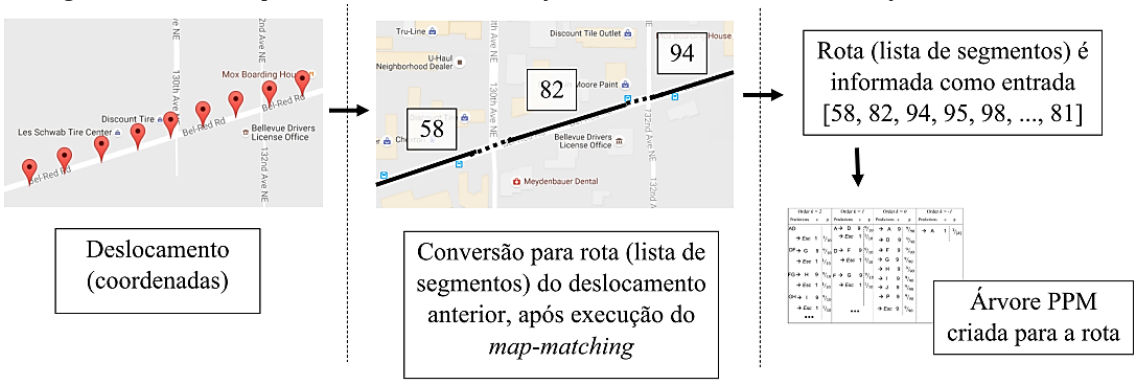
Ao utilizar o PPM, é possível recuperar todos os segmentos de uma rota, ou apenas parte dela. Em virtude desta característica, é possível prever o destino e a rota restante, com base (1) nos *segmentos já percorridos*, (2) na *origem do deslocamento* e (3) na *posição atual do usuário*. Com relação ao uso de cadeia de Markov e ao uso do HMM, seus usos se restringem à possibilidade de previsão apenas do destino, ao invés de toda a rota restante. Portanto, os modelos de previsão Markov e HMM, implementados nesta tese, são considerados como modelos de previsão preliminares, pois, para uma comparação mais aprofundada com os modelos baseados no PPM, seria necessário que os modelos Markov e HMM tivessem a capacidade de prever, além do destino, toda a rota restante. A decisão em se usar HMM desta maneira ocorreu em virtude deste estar categorizado na classe de problemas NP-difícil (BREJOVÁ, BROWN e VINAř, 2004) (ABE e WARMUTH, 1992), que pode gerar um alto custo computacional de processamento, quando uma instância do modelo de previsão estiver em execução. Ao combinar PPM com Markov ou com HMM, é possível prever, já no início do

deslocamento, o destino e a rota que o usuário deverá realizar, e corrigir essa previsão quando necessário, conforme o usuário vai se deslocando.

A Figura 7 apresenta uma instância do modelo de conversão de um deslocamento geográfico para uma árvore de símbolos PPM. Na porção mais à esquerda da figura, o deslocamento é representado por um conjunto de coordenadas geo-temporais, ordenado temporalmente. Na porção central da figura, foi aplicada uma técnica de *map-matching*, que resultou na associação das coordenadas geo-temporais aos segmentos. É possível perceber que esta técnica agrupa várias coordenadas em um único segmento, o que diminui a quantidade de dados a ser manuseado. Finalmente, na porção mais à direita da figura, a lista de segmentos (representada por um vetor de inteiros) é informada como entrada para o modelo e, com isso, são obtidas as árvores PPM que se referem aos deslocamentos.

Na etapa de aprendizagem dos modelos de previsão baseados em PPM, é criada uma árvore PPM para cada rota diferente que o usuário realiza. Como os modelos de previsão baseados em PPM são personalizados por usuário, mesmo que dois usuários diferentes tenham realizado uma mesma rota, duas árvores PPM serão criadas, sendo uma para cada usuário. O relacionamento entre uma árvore PPM e um usuário é de *um-para-um*, enquanto um único usuário pode ter (e geralmente tem) várias árvores PPM que representam suas respectivas rotas. De maneira formal, o parâmetro de entrada para a etapa de aprendizagem pode ser definido pela tripla *<u, R, L>*, em que *u* representa o usuário, *R* o conjunto de todos os deslocamentos geográfico e *L* o conjunto de regiões geográficas já visitadas pelo usuário.

Figura 7 - Cenários que ocorrem desde a obtenção do deslocamento até a obtenção da árvore PPM.



Fonte: Elaborada pelo autor.

Na etapa de previsão, todas as árvores PPM criadas na etapa de aprendizagem permanecem inalteradas, isto é, não há qualquer mudança na estrutura da árvore quando

a etapa de previsão de destino e rota é iniciada. De maneira formal, o parâmetro de entrada para a previsão pode ser definido pela tripla $<u, P, L>$, em que u representa o usuário, P o deslocamento geográfico (parcial) já realizado e L o conjunto de lugares já visitados pelo usuário. Os procedimentos para previsão ocorrem da seguinte maneira (para os três modelos baseados em PPM propostos):

1. Inicialmente, o modelo de previsão identifica a informação contextual da trajetória a ser prevista, como dia da semana, horário, lugar e papel do lugar da origem;
2. Em seguida, o modelo de previsão identifica todas as árvores PPM relacionadas ao histórico de deslocamentos realizados em contexto similar. Por exemplo, considere que uma rota a ser testada foi percorrida em uma segunda-feira, entre 7h e 8h da manhã, onde a origem foi uma região geográfica que se refere à *casa* do usuário. O modelo vai recuperar não somente rotas históricas que estão sob este exato contexto, mas também rotas que estão em contextos similares, como aquelas realizadas entre segunda e sexta, com um intervalo de uma hora para mais e para menos (as rotas realizadas entre 6h e 9h da manhã);
3. No próximo passo, o modelo de previsão identifica a rota a ser prevista, obtém as árvores de símbolos PPM recuperadas pelo passo anterior (2) e realiza a compressão da rota a ser prevista com cada árvore PPM. Cada compressão da rota a ser prevista com uma árvore PPM irá gerar uma *Razão de Compressão* (RC). Quanto maior for o valor da RC, mais similaridade a rota a ser prevista possui com a árvore PPM de uma rota já realizada;
4. Por último, para cada k árvore PPM, o modelo tem a capacidade de recuperar tanto a origem como o destino referente à rota representada pela árvore PPM. Após isso, o modelo de previsão funciona da maneira especificada abaixo:
    4.1. ***Se o modelo de previsão for PPM com Markov***, recupera-se a probabilidade de transição entre a origem i da rota a ser prevista e o destino j armazenado pela árvore PPM (probabilidade $p_{i,j}$). Consequentemente, para cada árvore PPM, são obtidos os valores RC_k e $p_{i,j}$, e estes dois valores são utilizados para o cálculo da métrica *ICP* (de *Increased Confidence Prediction*, em português *Aumento da Confiança de Previsão*) referente ao modelo PPM-Markov. Essa

métrica é obtida pela Equação (1), em que *k* se refere à *k-ésima* árvore PPM, ICP_k refere-se ao valor da métrica *ICP* para a *k-ésima* árvore PPM, RC_k refere-se ao valor de *RC* para a *k-ésima* árvore PPM e $p_{i,j}$ refere-se à probabilidade de sair da origem *i* para o destino *j*, que foi obtida pela *k-ésima* árvore PPM;

4.2. ***Se o modelo de previsão for PPM com HMM***, recupera-se a origem *i* da rota a ser testada juntamente com as observações (informações contextuais de dia da semana e horário), para obtenção da máxima probabilidade das observações (PO_{max}) para o modelo HMM $\lambda = (A, B)$ – também conhecido como o **Problema 2**. Assim, para cada árvore PPM, são obtidos os valores de referência RC_k e PO_{max}, e, estes dois valores, são utilizados para o cálculo da métrica *ICP* referente ao modelo PPM-HMM. Essa métrica é obtida pela Equação (2), em que *k* refere-se à *k-ésima* árvore PPM, ICP_k refere-se ao valor da métrica *ICP* para a *k-ésima* árvore PPM, RC_k refere-se ao valor de *RC* para a *k-ésima* árvore PPM e PO_{max} refere-se à probabilidade de ocorrência da sequência de estados mais provável, dada a sequência de observações.

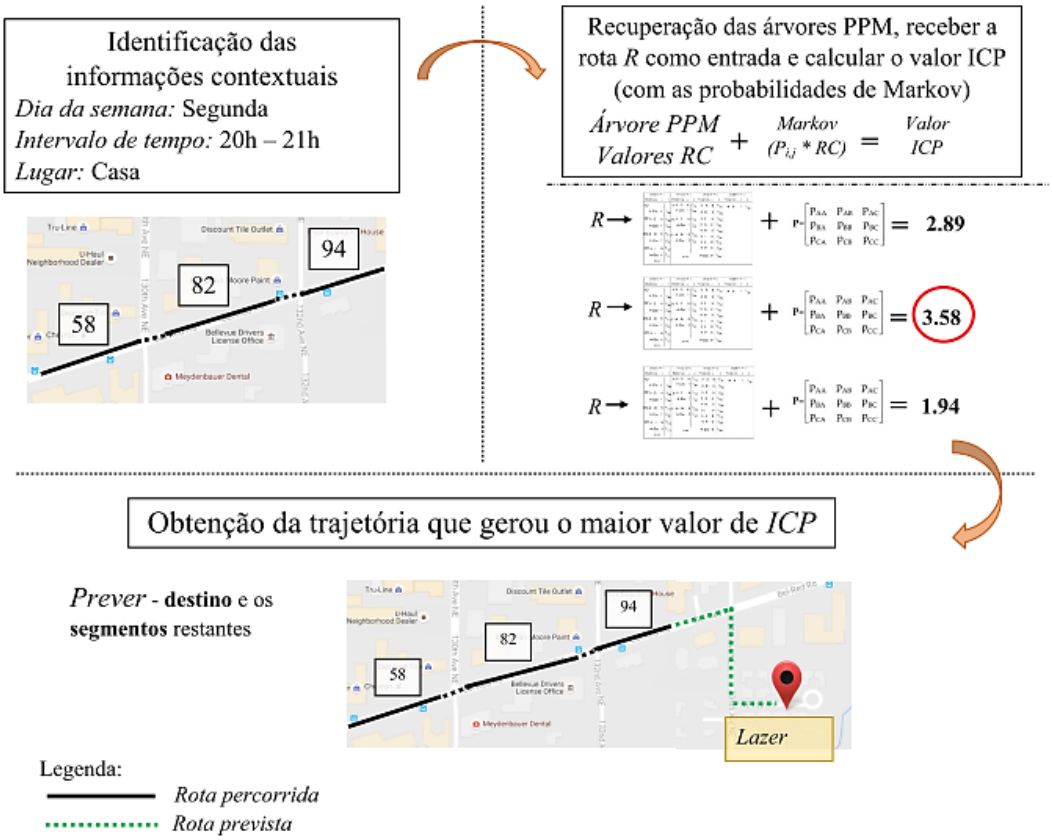
$$ICP_k = RC_k + (RC_k * p_{i,j}), \text{ para PPM-Markov} \quad \text{Equação (1)}$$

$$ICP_k = RC_k + (RC_k * PO_{max}), \text{ para PPM-HMM} \quad \text{Equação (2)}$$

Com relação à métrica ICP (elaborada na pesquisa descrita nesta tese), a faixa de valores que esta métrica pode receber varia de *RC* até *2 * RC*, em que *RC* é um número real positivo. Quanto às Equações (1) e (2), há uma ponderação do valor da *RC* pela probabilidade de transição (quando estiver combinada a uma cadeia de Markov) ou pela probabilidade de ocorrência da sequência de estados mais provável (quando for combinada a um HMM). O motivo da soma é que, mesmo que a probabilidade da cadeia de Markov ou do HMM resulte em zero (o que representa, conceitualmente, que o usuário nunca realizou certo trajeto), o valor de ICP poderá não resultar em zero. Com isso, embora um trajeto que tenha iniciado em uma origem totalmente nova, resultará em uma predição, pois o valor da *RC* contemplará as informações dos segmentos que estão sendo percorridos.

A Figura 8 apresenta uma instância do modelo de previsão PPM-Markov descrito textualmente, para a obtenção do valor ICP_k. Na parte mais à esquerda da figura, o passo inicial do preditor tem a responsabilidade de identificar as informações contextuais da rota a ser prevista (por exemplo, dia da semana e horário do deslocamento) e recuperar as árvores PPM sob condições semelhantes. Em seguida, na porção à direita da figura, é ilustrado o procedimento referente ao cálculo da métrica *ICP* (referente ao modelo PPM-Markov) para cada árvore PPM recuperada. Por último, na parte mais abaixo da figura, é identificada a árvore PPM que gerou o maior valor de *ICP*, a elege como sendo a prevista e recupera o destino e os segmentos restantes a partir da árvore.

Figura 8 - Cenários para a escolha do modelo de trajetória para ser fornecido como previsão, para o preditor que combina PPM e Markov.



Fonte: Elaborada pelo autor.

O modelo de previsão que combina PPM e HMM é muito semelhante ao da Figura 8. A diferença, no entanto, é que o procedimento de cálculo que está apresentado no meio da figura será diferente, uma vez que será utilizada a Equação (2) para o cálculo do ICP para PPM-HMM. Para o modelo de previsão que utiliza apenas PPM, a métrica *ICP* representará o mesmo valor da razão de compressão.

Os modelos de previsão baseados em PPM possuem um mecanismo para identificar se uma rota em curso nunca foi realizada pelo usuário, isto é, não foi utilizada na etapa de aprendizagem do modelo. Na próxima subseção, será explanado o funcionamento deste mecanismo.

4.2.1.1. Mecanismo para Previsão de um Lugar Nunca Visitado

Um grande desafio na área de previsão de trajetórias está no desenvolvimento de um modelo capaz de identificar que o usuário está percorrendo uma rota que nunca fora realizada, isto é, um modelo com a capacidade de discernir que o destino não será alcançado e a que rota inicialmente prevista não será feita pelo usuário. Este problema pode ser representado pelo seguinte questionamento: "*Em um deslocamento em curso, dada uma previsão, como o preditor pode identificar se o usuário vai percorrer os segmentos e alcançar o destino previsto?*". Para lidar com essa questão, os modelos de previsão propostos nesta tese implementam um mecanismo chamado de *rota reduzida*, que funciona da seguinte maneira:

1. Do deslocamento em curso a ser previsto (isto é, testado para obtenção das previsões), o modelo de previsão *reduz a rota* em um certo percentual;
2. A *rota reduzida* também será informada como parâmetro de entrada para a previsão, o que resultará em um valor da métrica *ICP*. Os valores da métrica *ICP* entre a rota original e a rota reduzida são comparados;
3. Se o valor da métrica *ICP* da rota original for maior que o valor da métrica *ICP* da rota reduzida, as previsões do destino e da rota inicialmente realizadas permanecem as mesmas. No entanto, se o valor da métrica *ICP* da rota reduzida for maior que o valor da métrica *ICP* da rota original, o preditor considera que o usuário está realizando uma rota totalmente nova;
4. O preditor recupera o **papel do lugar** do destino inicialmente previsto (para a rota parcial no tamanho original), mas desconsidera as previsões do lugar e da rota;
5. Após isso, o modelo busca o lugar mais próximo de onde o usuário está com o mesmo **papel do lugar** do destino inicialmente previsto, e sugere (prevê) este lugar como o destino possível que o usuário irá alcançar.

Ao realizar o procedimento descrito acima, o modelo tenta lidar com duas questões importantes relacionadas à previsão: 1) a possibilidade de identificação de que o usuário está realizando um trajeto totalmente novo e diferente do inicialmente previsto (em virtude da comparação do valor de *ICP* da rota original com o valor da rota reduzida); e 2) a capacidade de prever um destino onde o usuário nunca visitou. Com relação ao mecanismo de previsão de um destino nunca visitado, uma restrição é a não capacidade de prever corretamente um lugar que possua o papel de lugar diferente do papel de lugar do destino inicialmente previsto.

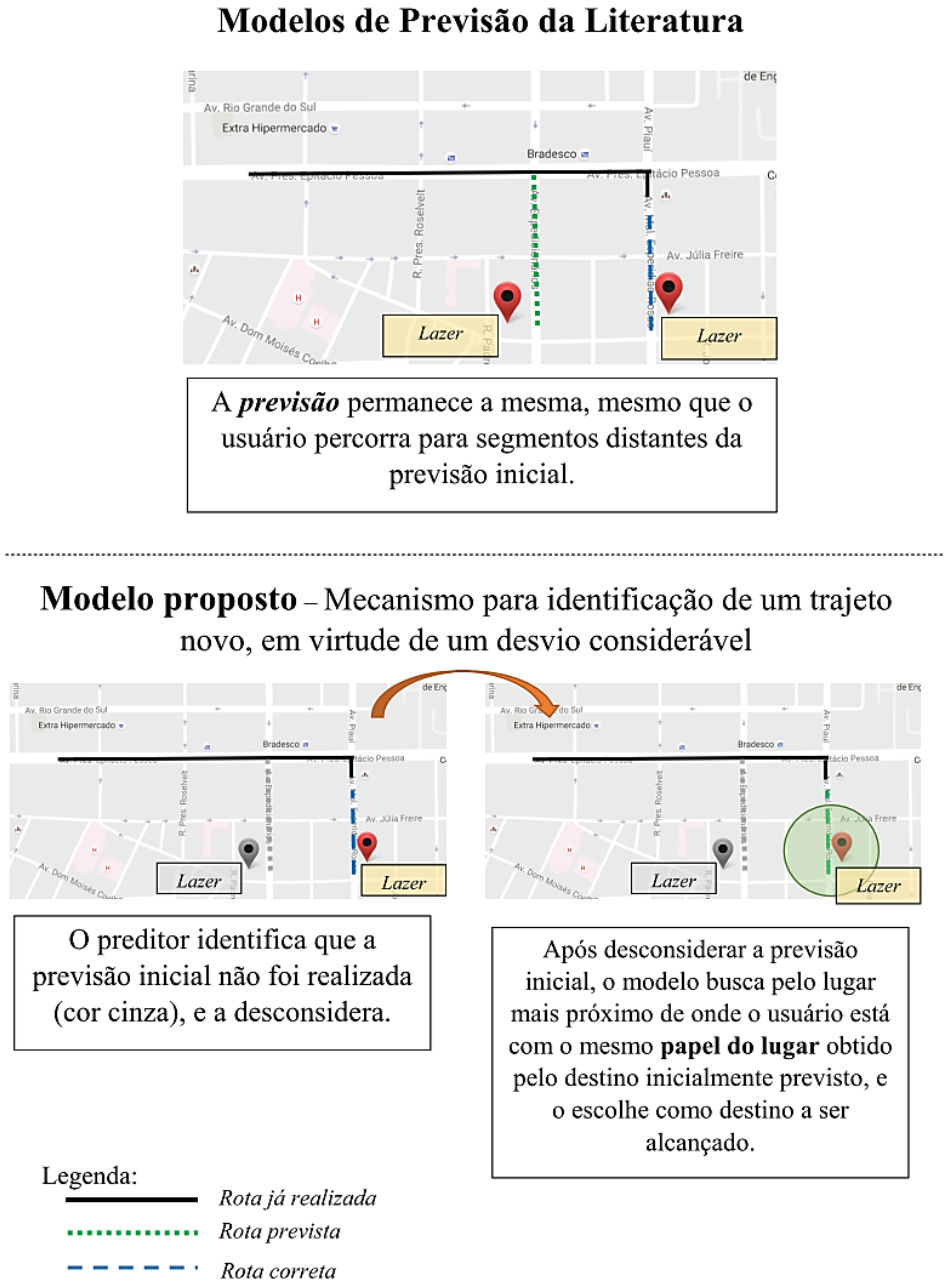
A Figura 9 ilustra uma instância do funcionamento do mecanismo para identificação de um lugar que o usuário nunca visitou. Em um modelo de previsão comum (na parte de cima da figura), é mantida a previsão de rota e destino mesmo que o usuário tenha percorrido uma rota bem diferente. Os modelos de previsão propostos nesta tese, na parte de baixo da figura, possuem a capacidade de identificar que o usuário tomou um desvio considerável da trajetória inicialmente prevista, e deve ir para um destino totalmente novo, por meio do mecanismo de *rota reduzida*. Assim, uma nova previsão de destino é realizada com base no papel do lugar do destino obtido pela previsão inicialmente realizada.

Para o funcionamento da previsão de um lugar nunca visitado, é imprescindível o uso do papel do lugar que uma região de parada representa para o usuário. Para isso, algumas regras de inferência do papel do lugar foram elaboradas como parte do modelo de previsão, e serão descritas na próxima subseção.

4.2.1.2. Inferência do Papel de um Lugar

Os modelos de previsão utilizam a informação de papel do lugar para poder prever um lugar nunca visitado pelo usuário. O papel do lugar, nesta pesquisa, foi obtido por meio de regras de inferências, isto é, sem receber prontamente a informação do papel do lugar. Para a inferência, foi elaborado um conjunto de regras, a partir de observação empírica dos dados conjuntamente com informações contextuais extraídas, tais como duração da permanência em um local, os dias e horários da semana que a visita costuma ocorrer e o tipo do POI obtido pelos serviços externos (*Google Places, Foursquare* e *Factual*). Foram considerados sete papéis diferentes de lugares: *casa, trabalho, lazer, esportes, educação, outros* e *desconhecido*.

Figura 9 - Funcionamento do mecanismo para identificar se o usuário está realizando um percurso novo.



Fonte: Elaborada pelo autor.

O Quadro 3 apresenta as regras definidas para cada um dos papéis de lugares considerados. Este quadro possui três colunas, sendo a primeira referente ao papel do lugar, a segunda que descreve a representação que aquele papel tem no contexto deste trabalho e a terceira é a regra de inferência. Para a definição da regra, foram utilizadas as informações de intervalo de tempo que o usuário permaneceu no lugar, os dias da semana que ele frequenta o lugar e o tipo do POI referente às consultas realizadas por um serviço externo (neste trabalho, *Google Place*, *Facebook* e *Factual*).

As regras de inferência disponíveis no Quadro 3 não conseguem contemplar todos os casos para inferência correta do papel do lugar e, portanto, estas definições estão sujeitas a modificações. Por exemplo, conforme as regras do quadro, não é possível inferir corretamente quando o usuário tiver dois lugares que representam o papel de lugar casa, e nem quando o usuário possuir um trabalho que não tenha expediente regular, como é o caso de uma pessoa que trabalhe como representante de vendas, ou um médico plantonista.

Quadro 3 - Regras de inferência para os tipos dos lugares.

Papel do lugar	**Representação**	**Regra**		
		Intervalo de tempo	**Dias da semana**	**Tipo do POI do serviço externo**
Casa	Lugar que o usuário mora. Este tipo só pode ocorrer uma única vez.	Mais de dez horas por dia.	Semana e final de semana.	Diferente de estabelecimento.
Trabalho	Lugar que o usuário trabalha. Pode ocorrer zero ou várias vezes.	Entre seis e oito horas por dia.	Semana e/ou final de semana.	Qualquer tipo.
Lazer	Lugares para lazer do usuário (por exemplo, restaurante ou cinema).	Mais de uma hora, desde que não seja sua *casa*.	Qualquer dia.	Restaurante, *shopping center*, cinema, hotéis e similares.
Esportes	Lugares para prática de esporte do usuário.	Entre uma e duas horas.	Qualquer dia.	Academia, ginásio, piscina e similares.
Educação	Lugares para estudo do usuário.	Entre duas e seis horas.	Entre segunda e sábado.	Escola, universidade, biblioteca e similares.
Outros	Lugares que o usuário visita para resolver tarefas.	Mais de 10 minutos, e que não sejam os tipos anteriores.	Qualquer dia.	Farmácia, padaria, clínica, banco, supermercado e similares.
Desconhe-cido	Lugares que não se enquadraram nos tipos anteriores.	-	Qualquer dia.	-

Fonte: Elaborada pelo autor.

4.2.2. Implementação dos Modelos de Previsão

Nesta seção, apresentam-se os principais algoritmos referentes aos modelos de previsão. O funcionamento dos algoritmos, referentes aos modelos de previsão, estão agrupados pelas etapas de aprendizagem e de treinamento.

4.2.2.1. Etapa de Aprendizagem

A etapa de aprendizagem consiste na construção do preditor para cada usuário. É neste momento que o perfil de deslocamento de um usuário é obtido, de forma personalizada e individual, sem considerar a influência de trajetos realizados por outros usuários. Será apresentado, inicialmente, o algoritmo para a aprendizagem da combinação de PPM com Markov. Em seguida, será descrita a diferença do algoritmo PPM com Markov para o algoritmo que utiliza a combinação de PPM com HMM, além de ser explicado o funcionamento do algoritmo que utiliza apenas PPM.

Algoritmo de Aprendizagem para PPM com Markov

O Algoritmo 1 detalha o funcionamento da etapa de aprendizagem da combinação do PPM com Markov, em que o algoritmo recebe como entrada um vetor com as rotas percorridas por um usuário (as rotas representadas por segmentos), o conjunto de pontos de paradas (já enriquecidas semanticamente) e o identificador do usuário (linhas 2 – 4). A saída gerada será o preditor personalizado criado ou atualizado (linha 6). Inicialmente, o algoritmo cria uma matriz de transição de estados, que representa as cadeias de Markov, ao utilizar as informações dos lugares visitados pelo usuário e das rotas para aprendizagem (linha 8). Os *estados* da cadeia de Markov, no contexto desta pesquisa, representam os *lugares visitados pelo usuário*. Após, será feita uma iteração sobre cada rota do vetor (linha 9), para obter os pontos de paradas inicial e final (isto é, o par <*origem*, *destino*> da rota) e a informação contextual (linhas 10 e 11). A rota, que está representada por um vetor de segmentos, é comprimida (linha 12), o que implica na geração de uma árvore de símbolos PPM, por meio da função *obter_árvore_ppm_da_rota(...)*. Na verdade, cada segmento possui um identificador único que o diferencia dos demais. Cada identificador representa um símbolo diferente e a união destes identificadores (que representam os segmentos das rotas) formarão uma palavra a ser comprimida pelo PPM, resultando em uma árvore de compressão. De posse das informações das paradas inicial

e final, informação contextual do deslocamento (como dia da semana da partida e horário de partida) e da árvore de símbolos PPM referente às rotas, será criada uma trajetória, que encapsula todas as informações referentes a um trajeto (linha 13). Por fim, a trajetória será inserida na variável referente ao conjunto de trajetórias (linha 14). Após a iteração sobre todas as rotas, o preditor é criado/atualizado com as informações referentes à matriz de probabilidades de Markov e às rotas representadas como árvores PPM, juntamente com as informações contextuais (linha 16).

Algoritmo 1 - Procedimento para modelagem do perfil de deslocamento de um usuário, usando PPM combinado com Cadeias de Markov.

```
 1:  ENTRADA
 2:    R   // Vetor com as rotas do usuário, representada por segmentos
 3:    E   // Conjunto de paradas com informação semântica
 4:    u   // Identificador do usuário
 5:  SAÍDA
 6:    P   // Modelo preditivo criado e/ou atualizado
 7:  MÉTODO
 8:    markov = criarCadeiaSimplesMarkov(R, E, u);
 9:    para cada rota existente em R faça {
10:      paradas = obter_paradas_inicial_e_final(rota, E);
11:      info = obter_info_contextual(rota);
12:      ppm-tree = obter_árvore_ppm_da_rota(rota);
13:      trajeto = criar_modelo_trajetoria(paradas, info, ppm-tree);
14:      armazenar_trajeto(trajeto, conjuntoDeTrajetos);
15:    } // Fim Para
16:    criar_ou_atualizar_preditor(P, markov, conjuntoDeTrajetos, u);
```

Fonte: Elaborada pelo autor.

Algoritmo de Aprendizagem para PPM com HMM

O algoritmo utilizado para a aprendizagem de PPM com HMM é semelhante ao Algoritmo 1. A diferença, porém, ocorre nas linhas 8 e 16 do Algoritmo 1, em que, no lugar de se construir uma matriz de probabilidade de transição de estados de Markov, será criado o HMM referente ao conjunto de observações. Assim, a linha 8 deste algoritmo, que possui o pseudocódigo

```
markov = criarCadeiaSimplesMarkov(R, E, u);,
```

deve ser substituída pelo pseudocódigo

```
hmm = criarHMM(R, E, u);,
```

em que a função *criarHMM(...)* cria uma única estrutura HMM para um usuário, e recebe as informações de todas as rotas percorridas pelo usuário, o conjunto de regiões de

paradas e o identificador do usuário. O pseudocódigo da linha 16 do Algoritmo 1 deverá ser reescrito para receber como parâmetro o HMM criado, ao invés da cadeia de Markov, resultando no pseudocódigo:

```
criar_ou_atualizar_preditor(P, hmm, conjuntoDeTrajetos, u);.
```

O conjunto de observações a ser utilizado para aprendizagem do HMM refere-se aos lugares de origem, aos dias da semana da partida e aos horários de partida das rotas a serem utilizadas para o treinamento. Com relação ao tamanho máximo de observações (TO) possíveis, este pode ser obtido pela fórmula,

$$TO = |L| + |D| + |H|$$

em que:

- $|L|$ representa a quantidade de lugares visitados pelo usuário (variável por usuário);
- $|D|$ representa a quantidade de dias da semana (sete dias); e
- $|H|$ representa a quantidade de hora (24 horas).

Com relação ao modelo HMM construído na etapa de aprendizagem, este pode ser representado por $\lambda = (A, B, \pi)$. Deste modelo, tem-se que:

- O *número de estados N* do modelo corresponde ao quantitativo de lugares visitados pelo usuário. Assim, os estados do HMM correspondem aos lugares visitados;
- O *número de símbolos de observações M* corresponde à combinação dos lugares de origem, com os dias da semana (sete dias) e com os horários (24 horas);
- A *matriz de transição de estados A*, de ordem $N \times N$, que contempla os lugares visitados pelo usuário;
- A *matriz de observação B*, de ordem $N \times M$, que contempla os lugares visitados e o as observações possíveis.

Algoritmo de Aprendizagem para PPM sem Combinação

O algoritmo utilizado para o uso do PPM sem combinação com outra técnica é similar ao Algoritmo 1. Porém, não é criado um objeto que represente nem uma cadeia de Markov (para o caso do algoritmo PPM-Markov) e nem um HMM (para o caso do

algoritmo PPM-HMM). No pseudocódigo apresentado no Algoritmo 1, isto é refletido pela remoção da linha 8. Uma vez que a linha 8 é removida do Algoritmo 1, os parâmetros reais informados para a ativação da função *criar_ou_atualizar_preditor(...)*, na linha 16, devem ser alterados, já que não existe um objeto correspondente ao objeto *Markov* e nem ao *HMM*, resultando no seguinte pseudocódigo:

```
criar_ou_atualizar_preditor(P, vazio, conjuntoDeTrajetos, u);,
```

em que o termo ***vazio*** representa a ausência na passagem de um objeto esperado como parâmetro da função.

4.2.2.2. Etapa de Previsão (Testes)

A etapa de previsão consiste na previsão de um destino e rota, dado um deslocamento em curso. Esta etapa descreve o funcionamento dos algoritmos utilizados para a previsão. Também é relevante destacar que essa previsão, além de realizada automaticamente, é feita em tempo real, isto é, à medida que o usuário se move, o destino a ser previsto pode ser recalculado.

Um teste, no contexto deste trabalho, é prever a região geográfica (destino) que o usuário deve ir, bem como a rota até esta região. Prever um destino que um usuário já visitou anteriormente pode ser mais fácil do que prever um lugar que ele nunca foi, uma vez que o deslocamento em curso pode ser similar a um deslocamento anterior. É pertinente destacar que as rotas usadas na etapa de testes são diferentes daquelas utilizadas para o treinamento.

Algoritmo de Previsão para PPM com Markov

O procedimento para execução da previsão, utilizando a combinação da técnica PPM com *Cadeias de Markov*, está detalhado no Algoritmo 2. O algoritmo recebe como entrada a rota parcial a ser prevista, o conjunto de regiões já geográficas visitadas pelo usuário e o identificador do usuário. Como saída, o algoritmo gera o maior valor *ICP* obtido e a trajetória prevista, com informações sobre o lugar de destino, a rota até alcançá-lo e o papel do lugar que o usuário deve ir.

Inicialmente, o algoritmo obtém a cadeia de Markov construída na etapa de aprendizagem (ou seja, a matriz de probabilidades de transição) do usuário (linha 10), e,

em seguida, recupera todo o conjunto de trajetórias já realizadas pelo usuário (linha 11). Então, é realizada uma iteração sobre o conjunto trajetórias obtidas do usuário em questão (linha 12), em que são obtidas as razões de compressão (RC) da rota em andamento (linha 13) e também são calculados os valores ICP (linha 14). As razões de compressão são obtidas por meio da função *obter_razão_compressão(...)*, e seu funcionamento ocorre da seguinte forma: (1) é recuperada a árvore PPM referente ao deslocamento de uma trajetória; (2) é feita a compressão da rota em curso (um conjunto de segmentos, isto é, um vetor de números inteiros) com a árvore PPM recuperada; e (3) é calculada a RC referente à compressão. O valor da métrica *ICP* obtida é armazenada juntamente com a trajetória (linha 15), para uso futuro. Por último, o algoritmo seleciona a trajetória que obteve o maior valor *ICP* dentre todos os valores calculados (linha 17), recupera o trajeto referente ao maior valor *ICP* obtido e o escolhe como previsto, tanto o destino como a rota para alcançar o destino.

Algoritmo 2 - Procedimento para teste das rotas, usando PPM combinado com as Cadeias de Markov.

```
 1:  ENTRADA
 2:    R      // Percurso que está sendo realizado
 3:    S      // Conjunto de pontos de paradas
 4:    u      // Identificador do usuário
 5:
 6:  SAÍDA
 7:    trajeto-previsto   // Trajetória prevista
 8:    maiorICP
 9:  MÉTODO
10:    markov = obter_cadeia_markov(u);
11:    M = obter_trajetórias(u);
12:    para cada trajeto existente em M faça {
13:      RC = calcular_razão_compressão(trajeto.árvorePPM, R);
14:      ICP = RC + (RC * markov[R.origem, trajeto.destino]);
15:      armazenar_valor_ICP(ICP, trajeto, valores-ICP);
16:    } // Fim Para
17:    maiorICP = obter_maior_ICP(valores-ICP);
18:    trajeto-previsto = prever_trajeto_restante(maiorICP, R);
19:   retornar trajeto-previsto
```

Fonte: Elaborada pelo autor.

Algoritmo de Previsão para PPM com HMM

O algoritmo de previsão referente à combinação do PPM com HMM é muito similar ao que é detalhado no Algoritmo 2, em que é recebido o mesmo conjunto de variáveis de entrada e gerada a mesma saída. A mudança, porém, ocorre apenas na linha

14 do Algoritmo 2. Enquanto a cadeia de Markov representa a possibilidade de transição entre os estados (no contexto deste trabalho, uma origem – já conhecida – e um destino provável) de um trajeto, o HMM representa a descoberta provável de um destino, dado um conjunto de observações. Portanto, almeja-se utilizar um HMM para a obtenção da probabilidade máxima, dado um conjunto de observações. Assim, a linha 14 do Algoritmo 2, que possui o pseudocódigo

```
ICP = RC + (RC * markov[R.origem, trajeto.destino]);,
```

deve ser substituída pelo pseudocódigo

```
hmmProb = obterMaiorProbalidadeHMM(R.origem,I.dia,I.intervalo);
ICP = RC + (RC * hmmProbability);,
```

em que a função *obterMaiorProbalidadeHMM (...)* recebe os parâmetros de origem atual do deslocamento, o dia da semana e o intervalo, e retorna a maior probabilidade obtida para o conjunto de observações informado. O conjunto de observações a ser informado ao HMM é representado pelo lugar de origem da rota em curso, o dia da partida e o horário da partida.

Algoritmo de Previsão para PPM sem Combinação

Quanto ao uso do PPM puro para previsão, sem combinação nem com Markov e nem com HMM, a linha 10, do Algoritmo 2, deve ser removida, e nenhum objeto referente às cadeias de Markov ou ao HMM deve ser recuperado. Esta remoção reflete em um ajuste que deve ser feito na linha 14, do Algoritmo 2, pois a variável *ICP* receberá, apenas, o valor da variável *RC* (Razão de Compressão), resultando no pseudocódigo *ICP = RC*.

Funcionamento dos Algoritmos de Previsão para o Mecanismo de Rota Reduzida

O Algoritmo 2 é executado duas vezes, para a previsão de uma única rota. A primeira execução utiliza a rota original a ser testada, enquanto que, na segunda vez, a rota a ser testada é a *rota reduzida,* que corresponde a 80% da rota original. Como explicado na Seção 4.2.1, os valores de *ICP* da rota original (100% da rota a ser testada) e da rota reduzida (80% da rota a ser testada) são comparados. Se o maior valor de ICP obtido for da rota original, o trajeto previsto inicialmente (destino e rota) permanece o mesmo. No entanto, se o maior valor de ICP for da rota reduzida, o preditor infere que o

usuário não vai ao destino inicialmente previsto, e identifica que o usuário está realizando uma rota nova. Com isso, o algoritmo busca o lugar mais próximo de onde o usuário está e o escolhe como novo destino provável, cujo tipo de lugar seja o mesmo do destino previsto inicialmente. Realizando este procedimento, o modelo de previsão proposto por este trabalho identifica um desvio feito pelo usuário de um trajeto previsto inicialmente e consegue prever um novo destino.

Testes com diferentes valores para o tamanho da *rota reduzida* sugerem que a maior efetividade deste mecanismo é obtida quando o parâmetro utilizado é de 0.8, isto é, quando a *rota reduzida* equivale a 80% da rota original.

4.3. Conclusão do Capítulo

Neste capítulo, foram apresentados os modelos de previsão de trajetórias propostos. Foram apresentados os quatro módulos principais responsáveis por receber os dados de um deslocamento de um usuário e prever seus locais de destino. No capítulo a seguir, serão apresentados os experimentos realizados com os modelos de previsão propostos, com suas respectivas análises estatísticas, úteis para avaliar e aperfeiçoar os modelos propostos.

Capítulo 5

Avaliação Experimental

Neste capítulo, são apresentados os experimentos referentes aos modelos de previsão propostos nesta pesquisa. Foi utilizada uma base de rotas pública para acesso e de usuários reais para avaliação (a base *MSR GPS Privacy Dataset*[2], disponibilizada pela Microsoft), fazendo com que os modelos avaliados lidem com situações e adversidades referentes a deslocamentos reais. Adicionalmente, serão apresentadas as análises estatísticas, conduzidas para avaliar se o procedimento adotado é válido estatisticamente. As análises estatísticas foram realizadas conforme procedimentos que podem ser consultados nos livros de Juristo e Moreno (JURISTO e MORENO, 2010), e de Montgomery e Runge (MONTGOMERY e RUNGER, 2011), e o detalhamento destas análises estão disponíveis no APÊNDICE A e no APÊNDICE B desta tese.

5.1. Seleção dos Dados

A base *MSR GPS Privacy Dataset* contém dados de 21 usuários que residem majoritariamente na cidade de Seattle, nos Estados Unidos, e teve uma duração de coleta de três meses. O conteúdo desta base possui as coordenadas espaciais (latitude e longitude) e horário em que os usuários estiveram nas coordenadas. Para a criação da base, foi utilizado um dispositivo de captura de GPS *Royaltek RBT-2300*, com captura de posicionamento a cada 5 segundos.

[2] Disponível em https://www.microsoft.com/en-us/download/details.aspx?id=54965

A base de dados MSR consiste em um grande arquivo de coordenadas geo-temporais para cada usuário. Ou seja, os arquivos não são separados por rota realizada (o que resultaria em muitos arquivos por usuário), ao invés disso, todas as localizações estavam em um único arquivo. Os módulos de *Identificação de paradas* e de *Obtenção de rotas* foram utilizados, portanto, para criação e segmentação das rotas por usuário, o que resultou em um total de mais de 1.500 rotas para todos usuários.

Com relação ao perfil da base MSR, muitos lugares foram visitados apenas uma vez, contribuindo, também, para que várias rotas tenham sido percorridas uma única vez. A Tabela 3 apresenta o perfil dos 10 usuários com mais rotas realizadas. A primeira coluna representa a ordenação da tabela conforme a quantidade de rotas percorridas por usuário. A segunda coluna identifica a numeração do usuário na base, enquanto a terceira coluna apresenta o quantitativo de lugares visitados pelo usuário. A quarta coluna apresenta o quantitativo de rotas realizadas apenas uma vez (subcoluna rotulada pelo número *1*) e duas ou mais vezes (subcoluna rotulada por *2+*). A quinta coluna lista os lugares visitados pelo tipo de lugar. É possível perceber que, mesmo com a execução do mecanismo de identificação de *tipos de lugares*, não foi possível identificar o local de trabalho para alguns casos, conforme as regras utilizadas para inferência de tipo de lugar adotadas nesta pesquisa. As subcolunas representam parte dos tipos de lugares considerados para inferência. A sexta, também a última, coluna da tabela apresenta o quantitativo de lugares visitados apenas uma vez (coluna *1*) e duas ou mais vezes (coluna *2+*).

Com relação aos 10 usuários com mais deslocamentos, tem-se que mais de 70% dos lugares foram visitados apenas uma vez e mais de 56% das rotas foram percorridas também uma única vez. Embora nenhum resultado estatístico tenha sido apresentado, é possível perceber, de forma intuitiva, que, prever trajetórias em bases de dados com muitas rotas realizadas apenas uma vez, apresenta uma dificuldade maior quando comparada com bases que possuem rotas rotineiras em sua composição majoritária. Desta forma, foram criados dois cenários para testes, que serão detalhados na próxima seção.

Tabela 3 - Cenário dos 10 usuários com mais rotas realizadas.

#	Usuário	Total Lugares	Rotas		Tipos de lugares				Lugares visitados	
			1	*2+*	*Casa*	*Trabalho*	*Lazer*	*Outro*	*1*	*2+*
1	18	72	51	60	1	0	14	58	54	18
2	19	82	52	57	1	0	58	23	58	24
3	5	44	29	79	1	0	31	13	29	15
4	3	69	48	56	1	1	38	20	56	13
5	8	46	51	53	1	1	38	7	24	22
6	16	82	77	25	1	0	63	18	60	22
7	21	81	70	29	1	0	48	32	57	24
8	20	109	62	19	1	0	66	42	71	38
9	15	85	56	18	1	1	47	37	60	25
10	4	90	44	25	1	0	58	32	68	22
Total		**760**	**540**	**421**	**10**	**3**	**461**	**282**	**537**	**223**

Fonte: Elaborada pelo autor.

5.2. Configuração dos Experimentos

Em virtude da grande quantidade de lugares visitados e rotas percorridas apenas uma vez, foram criados dois cenários de teste: 1) o cenário *TodosTrajetos*, que contempla todas as rotas da base MSR, isto é, não há qualquer tipo de filtragem, como, por exemplo, se a rota foi realizada uma ou duas vezes; e 2) o cenário *TrajetosMaiorQueDois*, que seleciona da base de rotas MSR apenas aquelas que foram percorridas duas vezes ou mais.

Com relação às métricas estatísticas utilizadas para avaliação dos modelos e para a resolução das Questões de Pesquisa, foram utilizadas *precisão, cobertura* e *medida-F* – isto é, a tradicional *F1 Score*, em que $\beta = 1$ (VAN RIJSBERGEN, 1979). As métricas utilizam as informações de *Verdadeiro Positivo* (TP), *Falso Positivo* (FP) e *Falso Negativo* (FN). O valor de TP indica a quantidade de previsões que foram feitas corretamente, enquanto o valor de FP representa a quantidade de previsões incorretas. O cálculo do valor de FN representa a quantidade de vezes que o destino correto a ser previsto não esteve entre os *N* possíveis lugares. Assim, para realizar uma determinada previsão, são retornados *N* destinos possíveis. Se dentre estes *N* destinos candidatos o

correto não estivesse presente nele, o valor de FN deve ser incrementado. Por exemplo, considere cinco previsões a serem realizadas. Se em duas delas o destino correto não foi recuperado, o valor de FN é igual a dois. As outras três previsões estariam distribuídas entre os valores de TP e FP. O valor de *Verdadeiro Negativo* (TN) não foi contemplado pelas métricas estatísticas utilizadas, porém, TN refere-se à quantidade de lugares não previstos e que, de fato, o usuário não chegou a ir. O valor de TN é utilizado para a obtenção da métrica de *taxa de acurácia*.

Nos experimentos, foi realizada validação cruzada, em que o conjunto de rotas foi particionado em 10 grupos (*10-fold cross-validation*), mutuamente exclusivos, isto é, cada grupo possui rotas diferentes das demais. Para treinamento, foram utilizados nove grupos (90% dos dados), enquanto o grupo restante foi utilizado para teste (10% dos dados). Os testes foram repetidos cinco vezes, na tentativa de obter resultados mais consistentes.

5.3. Resultados

Durante o planejamento dos testes, é importante definir o *design* de experimentos conforme o que se busca avaliar. O modelo teórico do *design* utilizado neste experimento foi o *fatorial* (com um único fator) *com blocagem* (MONTGOMERY e RUNGER, 2011). O fator e o bloco utilizados variam conforme a questão de pesquisa: por exemplo, para a questão de pesquisa 1 (**QP1**) o *fator* refere-se aos modelos de previsão, enquanto o *bloco* refere-se ao percentual de completude da rota a ser testada. O motivo de blocar a rota a ser testada é uma tentativa de homogeneizar o tamanho das rotas a serem testadas, e minimizar a influência do tamanho das rotas nos resultados. As variáveis resposta definidas para análise das questões de pesquisa, e suas respectivas hipóteses, são as métricas estatísticas *precisão*, *cobertura* e *medida-F*.

Foi realizada, em seguida, a análise de variância (ANOVA), para verificação do impacto do fator e da blocagem nos resultados obtidos, além de verificação da significância (ou confiança) estatística dos resultados obtidos pelo modelo, isto é, se o que foi obtido é confiável e pode ser obtido novamente em testes futuros. A ANOVA é aplicada quando os resultados obtidos entre os níveis do fator do experimento são diferentes entre si, sendo necessária a realização de uma análise estatística dos valores para comentar os testes. Para verificar a influência de um *fator F* e de um *bloco B* para

cada uma das variáveis resposta avaliadas (*precisão, cobertura* e *medida-F*), o modelo matemático utilizado é o que consta na Equação (1),

$$y_{ij} = \mu + \beta_i + \alpha_j + e_{ij} \quad \text{Equação (1)}$$

em que:

- y_{ij} representa a média da variável resposta sob análise obtida pela combinação do *i-ésimo* bloco com o *j-ésimo* fator;
- μ representa a média total da variável resposta sob análise;
- β_i representa o efeito do *i-ésimo* nível do *bloco B*;
- α_j representa o efeito da *j-ésima* alternativa do *fator F*; e
- e_{ij} representa o erro experimental (ou residual) da combinação do *i-ésimo* bloco com o *j-ésimo* fator.

Os dados apropriados para a análise de variância devem possuir, além das médias das variáveis resposta (*precisão, cobertura* e *medida-F*), os valores referentes aos *efeitos* gerados pelos níveis do bloco e do fator.

Para realizar a análise de variância e descobrir a influência do fator, da blocagem e dos erros, devem ser calculados a soma dos quadrados (SQ) da variável resposta sob análise (SSY), a SQ referente ao fator investigado (SSA), a SQ da média geral da métrica sob análise (SS0), a SQ referente ao bloco (SSB), a SQ dos erros (SSE) e a SQ do resíduo geral (SST).

A próxima etapa é realizar o teste de significância, ou teste de hipóteses. Nesta análise, será aplicado o *Teste F*. Inicialmente, devem ser obtidos o cálculo de F e comparado este valor com o valor crítico de (F_c) disponível na tabela F, considerando a significância adotada, e os graus de liberdade das alternativas e do resíduo. Para calcular o valor de F da análise de variância do experimento, divide-se a média quadrática do fator (MSA) pela média quadrática dos erros (MSE). O MSA é obtido pela divisão do SSA pelo grau de liberdade (GL) do fator menos um, conforme Equação (2).

$$MSA = \frac{SSA}{GL(fator) - 1} \quad \text{Equação (2)}$$

Já o MSE é obtido com a divisão do SSE pela multiplicação dos graus de liberdade do fator menos um com o GL do bloco menos um, conforme Equação (3).

$$MSE = \frac{SSE}{[GL(fator) - 1] * [GL(bloco) - 1]} \quad \text{Equação (3)}$$

Por último, divide-se o valor de MSA por MSE e obtém-se o valor de F para o experimento. Este valor de F deve ser comparado com o valor F_c da tabela F, considerando os graus de liberdade do experimento em questão e também a confiança estatística que se quer adotar (como, por exemplo, 95% de confiança). Caso o valor de F seja maior que o valor F_c, tem-se a possibilidade de refutar a hipótese nula que foi adotada para a questão de pesquisa que se quer responder. Refutar a hipótese nula significa descartar, sob determinada confiança estatística, a hipótese de que os resultados das médias obtidos pelas alternativas do *fator F* são iguais.

Para o uso da ANOVA, porém, deve-se verificar se os dados obtidos atendem ao requisito de serem oriundos de uma distribuição normal. Esta verificação deve ser realizada com a análise dos erros experimentais (ou residuais), em que deve ser elaborado um gráfico quantil-quantil (ou gráfico Q-Q, do inglês *Q-Q plot*), além da aplicação do teste de normalidade *Shapiro-Wilk*, uma vez que este teste é o mais eficaz para verificação da normalidade dos dados (RAZALI, 2011). Quanto à aplicação do teste de normalidade *Shapiro-Wilk*, caso o *valor-p* (do inglês *p-value*) obtido seja maior que o *nível da significância* (isto é, *valor-p > nível da significância*), *significa que não é possível refutar a hipótese nula (H_0), em que a hipótese nula especifica que os dados são oriundos de uma população normal*.

Com relação ao uso do nível de confiança para avaliação experimental, os valores adotados costumam ser de 90%, 95% e 99% (MONTGOMERY e RUNGER, 2011), sendo o valor de 95% o mais comumente utilizado (COWLES e DAVIS, 1982). Portanto, nesta pesquisa, é utilizado o nível de confiança de 95%, tanto para avaliação dos resultados obtidos pelo teste ANOVA, como também para os resultados da avaliação do teste de normalidade.

Para melhor legibilidade dos experimentos, esta seção se subdivide nas análises estatísticas conforme as questões de pesquisas formuladas na Seção 1.2. As subseções estão organizadas para apresentar os resultados separadamente, conforme as variáveis resposta (isto é, as métricas estatísticas de *precisão*, *cobertura* e *medida-F)* que estão sendo analisadas.

5.3.1. Questão de Pesquisa – Comparação dos Modelos de Previsão

Esta seção apresenta, de maneira sumarizada, a análise estatística dos resultados para responder às hipóteses formuladas pela questão de pesquisa ***QP1*** apresentada na Seção 1.2, que está textualmente descrita abaixo.

QP1 *– Existe diferença no resultado do uso dos modelos de previsão de rotas e destino (implementados neste trabalho), incluindo previsão de lugares nunca visitados, com relação às métricas estatísticas (precisão, cobertura e medida-F) utilizadas para avaliação?*

H1-0: Não há influência entre os resultados obtidos, com relação ao uso dos modelos de previsão implementados, referentes às métricas estatísticas utilizadas para avaliação.

H1-1: Há influência entre os resultados obtidos, com relação ao uso dos modelos de previsão implementados, referentes às métricas estatísticas utilizadas para avaliação.

No procedimento de testes, cada rota original a ser testada gera três novas rotas, em que uma representa 15% da rota original, outra 50% e outra 85%. No planejamento do experimento para a resolução da **QP1**, foi adotado o *design fatorial* (com um único fator) *com blocagem*. O *fator F* refere-se aos *modelos de previsão*, enquanto o *bloco B* representa o *percentual de completude das rotas a serem testadas*. O *fator F* possui cinco níveis (que são os preditores), a saber: PPM-Markov; PPM-HMM; PPM; Markov e HMM. É importante ressaltar que o modelo de previsão que utiliza cadeia de Markov e o que utiliza HMM foram desenvolvidos para prever apenas o próximo destino, mas não preveem os segmentos para alcançar o destino previsto. Portanto, os resultados obtidos nesta avaliação experimental, principalmente para o HMM, ainda são preliminares e poderão ser investigados com maior aprofundamento, em um momento posterior a esta tese. Já *bloco B* possui três níveis, que são 15%, 50% e 85% referente à completude da rota a ser testada. Os testes foram realizados separadamente para cada um dos dois cenários de teste, um para o cenário (1) *TodosTrajetos* e outro para o cenário (2) *TrajetosMaiorQueDois*.

Todos os resultados apresentados nas subseções seguintes utilizam, referentes à **QP1**, o teste de *Análise de Variância* (ANOVA). Além disso, também é aplicado um teste para verificação de normalidade dos dados, isto é, para verificar se os dados utilizados para análise são oriundos de uma distribuição normal. Todos os resultados e análises, referentes à Questão de Pesquisa 1, estão detalhados no APÊNDICE A deste documento.

5.3.1.1. Testes com Cenário (1) *TodosTrajetos*

Nesta subseção, serão apresentados os resultados dos testes da **QP1**, sob a perspectiva da base de dados do cenário (1) *TodosTrajetos*. As variáveis resposta analisadas pelos testes foram *precisão, cobertura* e *medida-F*. Todos os testes foram repetidos cinco vezes, assim, as tabelas com os resultados das métricas analisadas apresentarão as médias obtidas após os cinco testes.

Nas Tabelas 4, 5 e 6 são apresentados, respectivamente, os resultados obtidos para as métricas de *precisão*, de *cobertura* e de *medida-F*. Na primeira coluna das tabelas, é especificada a variável de blocagem, que possui três níveis (15%, 50% e 85% da rota percorrida). A segunda coluna das tabelas contempla a variável referente ao fator de análise, que são os modelos de previsão. Nesta segunda coluna, há uma subdivisão em outras cinco colunas, referentes aos níveis do fator analisado. A última coluna apresenta a média obtida referente à métrica em análise, para cada nível do bloco, e a última linha apresenta a média da métrica obtida, para cada nível do fator.

Tabela 4 **–** Resultados obtidos para a métrica *Precisão* – Fator (Preditores) x Bloco (Percentual da rota percorrida), para o cenário de teste (1) TodosTrajetos.

Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,63	0,44	0,47	0,23	0,14	0,38
50%	0,67	0,50	0,54	0,23	0,14	0,42
85%	0,66	0,54	0,64	0,23	0,14	0,44
Média	**0,65**	**0,49**	**0,55**	**0,23**	**0,14**	**0,41**

Fonte: Elaborada pelo autor.

Tabela 5 - Resultados obtidos para a métrica *Cobertura* – Fator (Preditores) x Bloco (Percentual da rota percorrida), para o cenário de teste (1) TodosTrajetos.

Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,32	0,24	0,36	0,30	0,51	0,35
50%	0,38	0,28	0,48	0,30	0,51	0,39
85%	0,44	0,39	0,63	0,30	0,51	0,45
Média	**0,38**	**0,30**	**0,49**	**0,30**	**0,51**	**0,40**

Fonte: Elaborada pelo autor.

Tabela 6 - Resultados obtidos para a métrica *Medida-F* – Fator (Preditores) x Bloco (Percentual da rota percorrida), para o cenário de teste (1) TodosTrajetos.

Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,42	0,31	0,41	0,23	0,21	0,32
50%	0,49	0,36	0,51	0,23	0,21	0,36
85%	0,53	0,39	0,64	0,23	0,21	0,40
Média	**0,48**	**0,35**	**0,52**	**0,23**	**0,21**	**0,36**

Fonte: Elaborada pelo autor.

Análise Subjetiva dos Resultados Obtidos

Após a aplicação do teste ANOVA, no contexto do cenário de testes (1) (com a base) *TodosTrajetos*, foi possível refutar a hipótese nula da **QP1**, de que não há diferença entre os modelos de previsão avaliados, para as métricas de *precisão* e de *medida-F*. Para a variável resposta de *cobertura*, porém, não foi possível refutar a hipótese nula. Quanto à verificação da normalidade, não foi possível refutar a hipótese nula, isto é, não foi possível refutar que os dados obtidos são oriundos de uma distribuição normal. Assim, dando continuidade à análise, com uma linha mais subjetiva, pôde-se obter as seguintes considerações:

- Os valores para *precisão* foram maiores ao combinar PPM com Markov, seguido pelo uso do PPM puramente. Quando 15% da rota é percorrida, PPM-Markov demonstrou uma *precisão* de 0.63 comparado a 0.47 do PPM, uma diferença de mais 20% a mais para o primeiro tipo de preditor. Já quando a rota está próxima de ser concluída (85% de completude da rota), PPM-Markov continuou superando o PPM, obtendo-se, respectivamente, os valores 0.66 e 0.64. Porém, a diferença entre os preditores foi de menos de 5%. Pelos resultados dos testes, tem-se que 96,5% dos resultados foram influenciados pelo *fator F*. Assim, a combinação de métodos de previsão foi uma decisão coerente, estatisticamente adequada e pode ser um novo caminho de investigação na área de previsão de trajetórias, principalmente quando a base de dados possui trajetórias realizadas apenas uma vez. Outra observação que pode ser feita é que os modelos de previsão baseados no PPM obtiveram valores maiores para a *precisão*, o que sugere que a capacidade do PPM de se moldar ao deslocamento do usuário, à medida que o deslocamento vai ocorrendo, melhora a possibilidade da precisão;

- Sob 95% de nível de confiança, não foi possível refutar a hipótese nula para a variável resposta *cobertura*. Porém, se fosse adotado o nível de confiança de 90%, seria possível refutar a hipótese nula. A influência do *fator* (modelos de previsão) na *cobertura* foi consideravelmente menor do que na *precisão*, com apenas 68,2% de importância na obtenção dos resultados;
- Embora tenha sido possível refutar a hipótese nula (**H1-0**) para a variável resposta *precisão*, mas não tenha sido possível refutar a **H1-0** para a variável resposta *cobertura*, é importante avaliar o comportamento estatístico para a variável *medida-F*, que representa a média harmônica da *precisão* e da *cobertura*. Para a *medida-F*, assim como ocorreu com a variável *precisão*, foi possível refutar a **H1-0**. Os maiores valores obtidos para a *medida-F* foram para os modelos de previsão com base no PPM. Para 15% da rota percorrida, os modelos de previsão PPM-Markov e PPM puramente obtiveram valores muito próximos. Para 85% da rota percorrida, porém, o modelo de previsão PPM obteve um valor para a *medida-F* de, aproximadamente, 20% maior do que o modelo de previsão PPM-Markov. Assim, mediante avaliação dos resultados, os modelos de previsão com base em PPM apresentaram uma eficácia maior, para a previsão de trajetórias, comparados aos modelos Markov e HMM, com respeito às métricas de *precisão* e de *medida-F*.

5.3.1.2. Testes com Cenário (2) *TrajetosMaiorQueDois*

Os resultados desta subseção estão relacionados com o cenário de testes (2) *TrajetosMaiorQueDois*, em que a base de deslocamentos foi filtrada, sendo selecionadas apenas as rotas cujo par <*origem*, *destino*> foi realizado duas ou mais vezes. As variáveis resposta analisadas pelos testes também foram *precisão*, *cobertura* e *medida-F*, e estão apresentadas, respectivamente, nas Tabelas 7, 8 e 9. Todos os testes foram repetidos cinco vezes, assim, as tabelas com os resultados das métricas analisadas apresentarão as médias obtidas após os cinco testes.

Com relação às Tabelas 7, 8 e 9, na primeira coluna, é especificada a variável de blocagem, que possui três níveis (15%, 50% e 85% da rota percorrida). Na segunda coluna, está a variável referente ao fator de análise, que são os modelos de previsão, e contempla cinco modelos para análise. A última coluna das tabelas apresenta a média

obtida referente à métrica em análise, para cada nível do bloco, e a última linha apresenta a média da métrica obtida, para cada nível do fator.

Tabela 7 - Resultados obtidos para a métrica *Precisão* – Fator (Preditores) x Bloco (Percentual da rota percorrida), para o cenário de teste (2) TrajetosMaiorQueDois.

Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,77	0,50	0,62	0,70	0,19	0,56
50%	0,80	0,58	0,67	0,70	0,19	0,59
85%	0,82	0,61	0,74	0,70	0,19	0,61
Média	**0,80**	**0,56**	**0,68**	**0,70**	**0,19**	**0,59**

Fonte: Elaborada pelo autor.

Tabela 8 - Resultados obtidos para a métrica Cobertura – Fator (Preditores) x Bloco (Percentual da rota percorrida), para o cenário de teste (2) TrajetosMaiorQueDois.

Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,86	0,27	0,63	0,80	0,41	0,59
50%	0,89	0,39	0,73	0,80	0,41	0,64
85%	0,91	0,52	0,79	0,80	0,41	0,69
Média	0,89	0,39	0,72	0,80	0,41	**0,64**

Fonte: Elaborada pelo autor.

Tabela 9 - Resultados obtidos para a métrica Medida-F – Fator (Preditores) x Bloco (Percentual da rota percorrida), para o cenário de teste (2) TrajetosMaiorQueDois.

Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,81	0,35	0,63	0,71	0,24	0,55
50%	0,85	0,47	0,7	0,71	0,24	0,59
85%	0,87	0,56	0,76	0,71	0,24	0,63
Média	0,84	0,46	0,70	0,71	0,24	**0,59**

Fonte: Elaborada pelo autor.

Análise dos Resultados Obtidos

Conforme aplicação do teste ANOVA para o cenário de testes (2), que contempla o cenário (com a base) *TrajetosMaiorQueDois*, foi possível refutar a hipótese nula da **QP1**, de que não há diferença entre os modelos avaliados em termos de *precisão*, *cobertura* e *medida-F*. Além disso, de maneira semelhante ao cenário de testes (1), não foi possível refutar a hipótese nula referente à normalidade dos dados utilizados no teste ANOVA. Assim, dando continuidade à análise, com uma linha mais subjetiva, pôde-se obter as seguintes considerações:

- Quanto à variável resposta de *precisão*, o modelo de previsão PPM-Markov foi superior comparado aos demais, para todos os percentuais de completude da rota, resultando no valor 0,82. Para os valores de 15% e 50% da rota percorrida, o modelo de previsão PPM-Markov foi superior ao modelo que usa Markov em 10% e, aproximadamente, 20% superior ao modelo PPM. Com 85% da rota percorrida, o modelo de previsão PPM-Markov foi superior em quase 20% ao modelo de previsão de Markov, e 10% superior ao modelo PPM puramente. É importante destacar que, para o cenário de teste (1), os modelos de previsão PPM-Markov, PPM-HMM e PPM puramente proporcionaram a obtenção de uma *precisão* maior, quando comparado aos modelos Markov e HMM. Para o cenário de teste (2), os resultados obtidos pelo preditor PPM-Markov foram superiores aos resultados dos modelos Markov e HMM para todos os níveis do bloco. No cenário de testes (2), assim como no cenário (1), o *fator F* influenciou mais nos resultados obtidos (98%), quando comparado com o bloco e com os erros experimentais, significando que, com evidência estatística, blocar o experimento pelo percentual de completude da rota foi uma decisão correta. De forma sumarizada, para ambos os cenários, há evidências para afirmar que os modelos de previsão PPM-Markov e PPM puramente, para a base de dados MSR, possuem uma *precisão* superior comparado aos modelos Markov e HMM;
- Com relação à métrica de *cobertura* para o cenário de testes (2), foi possível refutar a hipótese nula **H1-0**, sob um nível de confiança de 95%, o que não ocorreu com o cenário de testes (1). O modelo PPM-Markov alcançou um valor de cobertura de 0,91, quando 85% da rota for percorrida. Comparando com os demais modelos, PPM-Markov obteve uma *cobertura* quase 15% maior do que o segundo maior valor, que foi obtido com Markov (0,80) e com PPM puramente (0,79). Quanto à influência, foi obtido que o *fator F* influenciou em 93% nos resultados para o cenário de testes (2).

Em face da implementação dos modelos de previsão Markov e HMM restringirem-se à previsão do destino, faltando-lhes a previsão da rota restante, faz-se

necessária uma investigação mais profunda acerca de como remediar esta limitação, estando esta fora do escopo desta tese.

5.3.2. Questão de Pesquisa – Influência da Base de Dados

Esta seção apresenta, de maneira sumarizada, o resultado da análise estatística dos dados para responder às hipóteses formuladas pela Questão de Pesquisa 2 (***QP2***) apresentada na Seção 1.2, que está textualmente descrita abaixo.

QP2 – *Existe diferença no resultado da previsão de trajetórias em uma base de dados com rotas que foram realizadas mais frequentemente (isto é, onde o par <origem, destino> foi realizado pelo menos duas vezes) versus uma base de dados que possui mais rotas que foram realizadas uma única vez (isto é, onde o par <origem, destino> foi realizado apenas uma vez) para os modelos de previsão baseados em PPM?*

H2-0: Não há diferença no resultado da previsão de trajetórias em usar bases com rotas frequentes *versus* bases com rotas que contenham muitos trajetos realizados apenas uma vez para previsão de trajetória, referentes às métricas estatísticas *precisão, cobertura* e *medida-F*.

H2-1: Há diferença no resultado da previsão de trajetórias em usar bases com rotas frequentes *versus* bases com rotas que contenham muitos trajetos realizados apenas uma vez para previsão de trajetória, referentes às métricas estatísticas *precisão, cobertura* e *medida-F*.

O motivo desta questão de pesquisa é aprofundar a investigação sobre os modelos com base em PPM, que são os modelos inovadores propostos nesta tese, realizando uma avaliação referente às métricas *precisão, cobertura* e *medida-F*. Além disso, conforme análise da QP1, os modelos de previsão com base no PPM apresentaram resultados superiores ou próximos às cadeias de Markov e ao HMM.

O procedimento de testes para analisar as hipóteses referentes a esta questão de pesquisa possuirá investigação diferente da realizada para a ***QP1***. Neste momento, o foco é verificar se a influência do fator *uso da base de dados* na previsão de trajetórias. Novamente, o *design* de experimento para os testes desta análise utilizado foi o *fator com blocagem*, em que o *fator F* refere-se ao *uso da base de dados*, e o *bloco B* refere-se ao *percentual de completude da rota a ser testada*.

Os níveis do *fator F* são: cenário (1) *TodosTrajetos*, que contém todas as rotas da base de deslocamentos utilizadas para avaliação, em que, nessa base, mais de 70% das rotas foram realizadas apenas uma vez; e cenário (2) *TrajetosMaiorQueDois*, que é composta pela seleção de rotas cujo par <*origem, destino*> foi realizado pelo menos duas vezes. Os níveis do *bloco B* são os mesmos da **QP1**, a saber: 15% da rota a ser testada; 50% da rota; e 85% da rota para teste. Os experimentos, nesta questão de pesquisa, foram realizados separadamente para cada modelo de previsão baseado em PPM, isto é, os resultados estão agrupados conforme os modelos PPM-Markov, PPM-HMM e PPM.

Todos os resultados apresentados nas subseções seguintes utilizam, referentes à **QP2**, o teste de *Análise de Variância* (ANOVA). Além disso, também é aplicado um teste para verificação de normalidade dos dados, isto é, para verificar se os dados utilizados para análise são oriundos de uma distribuição normal. Todos os resultados e análises, referentes à Questão de Pesquisa 2, estão mais detalhados no APÊNDICE B deste documento.

5.3.2.1. Testes com o Modelo de Previsão PPM-Markov

Esta seção apresenta os experimentos agrupados pelo modelo de previsão PPM com Markov. Para o experimento, o *fator F* possui os níveis referentes ao uso da base de dados, cujos valores podem ser a base *TodosTrajetos* e a base *TrajetosMaiorQueDois*, enquanto o *bloco B* refere-se ao percentual de completude da rota a ser testada (15%, 50% e 85%). As variáveis resposta analisadas foram *precisão*, *cobertura* e *medida-F*.

As Tabelas 10, 11 e 12 apresentam, respectivamente, os resultados obtidos para as métricas de *precisão*, *cobertura* e *medida-F*, após a realização do experimento. A primeira coluna das tabelas é destinada aos níveis do *bloco B*, enquanto a segunda coluna apresenta os resultados das médias obtidos para a métrica em análise, referente aos dois níveis do *fator F*. A última coluna apresenta as médias obtidas por bloco, e a última linha apresenta as médias de cada uma das duas alternativas do *fator F*.

Tabela 10 - Resultados obtidos para a métrica de *Precisão* – Fator (Base de dados) x Bloco (Percentual de completude da rota), para o modelo PPM-Markov.

Bloco B	***Fator F* – Base de dados**		**Média**
	TodosTrajetos	*TrajetosMaiorQueDois*	
15%	0.63	0.77	0.70
50%	0.67	0.80	0.74
85%	0.66	0.82	0.74
Média	**0.65**	**0.80**	**0.73**

Fonte: Elaborada pelo autor.

Tabela 11 - Resultados obtidos para a métrica de Cobertura – Fator (Base de dados) x Bloco (Percentual de completude da rota), para o modelo PPM-Markov.

Bloco B	***Fator F* – Base de dados**		**Média**
	TodosTrajetos	*TrajetosMaiorQueDois*	
15%	0.32	0.86	0.59
50%	0.38	0.89	0.64
85%	0.44	0.91	0.68
Média	**0.39**	**0.89**	**0.63**

Fonte: Elaborada pelo autor.

Tabela 12 - Resultados obtidos para a métrica de Medida-F – Fator (Base de dados) x Bloco (Percentual de completude da rota), para o modelo PPM-Markov.

Bloco B	***Fator F* – Base de dados**		**Média**
	TodosTrajetos	*TrajetosMaiorQueDois*	
15%	0.42	0.81	0.62
50%	0.49	0.85	0.67
85%	0.53	0.87	0.70
Média	**0.48**	**0.84**	**0.66**

Fonte: Elaborada pelo autor.

Análise dos Resultados Obtidos

Conforme os resultados obtidos para as variáveis resposta (métricas estatísticas) de *precisão*, *cobertura* e *medida-F*, é possível refutar a hipótese nula (H2-0) da questão de pesquisa 2 (**QP2**), e afirmar que não há evidência estatística de que as médias dos valores obtidos para as bases de dados *TodosTrajetos* e *TrajetosMaiorQueDois* são iguais, ao variar o fator em análise. Além disso, o fator (base de rotas) demonstrou ser a informação mais influente nos resultados obtidos (com índices de influência de 98%) do que o bloco ou mesmo outro resíduo do experimento.

Após a análise dos resultados estatísticos, é possível verificar que o fator (base de dados) obteve elevada influência na obtenção dos resultados. Isto é, utilizar uma base de

dados que é majoritariamente composta por rotas realizadas apenas uma vez (base *TodosTrajetos*) é mais difícil de ser classificada comparada a bases que tenham rotas que foram realizadas duas ou meias vezes (base *TrajetosMaiorQueDois*). Com esse cenário de resultado da **QP2**, observa-se que o modelo de previsão PPM-Markov é adequado como modelo de previsão, uma vez que na **QP1** (quando foram avaliados os modelos de previsão) também foi identificado que ele possuiu influência na obtenção dos resultados de previsão. Assim, como os resultados das métricas foram bons e apresentaram significância estatística, o preditor PPM-Markov aparenta ser apropriado para previsão de rotas e destinos.

5.3.2.2. Testes com o Modelo de Previsão PPM-HMM

Esta seção apresenta os experimentos agrupados pelo modelo de previsão PPM com HMM. As Tabelas 13, 14 e 15 apresentam, respectivamente, os resultados obtidos para as métricas de *precisão*, *cobertura* e *medida-F*, após a realização do experimento. A primeira coluna das tabelas é destinada aos níveis do *bloco B*, enquanto a segunda coluna apresenta os resultados das médias obtidos para a métrica em análise, referente aos dois níveis do *fator F*. A última coluna apresenta as médias obtidas por bloco, e a última linha apresenta as médias de cada uma das duas alternativas do fator.

Tabela 13 - Resultados obtidos para a métrica de *Precisão* – Fator (Base de dados) x Bloco (Percentual de completude da rota), para o modelo PPM-HMM.

Bloco B	***Fator F* – Base de dados**		**Média**
	TodosTrajetos	*TrajetosMaiorQueDois*	
15%	0,44	0,5	0,47
50%	0,50	0,58	0,58
85%	0,54	0,61	0,61
Média	**0,49**	**0,56**	**0,67**

Fonte: Elaborada pelo autor.

Tabela 14 - Resultados obtidos para a métrica de *Cobertura* – Fator (Base de dados) x Bloco (Percentual de completude da rota), para o modelo PPM-HMM.

Bloco B	***Fator F* – Base de dados**		**Média**
	TodosTrajetos	*TrajetosMaiorQueDois*	
15%	0,24	0,27	0,26
50%	0,28	0,39	0,34
85%	0,39	0,52	0,46
Média	**0,30**	**0,39**	**0,35**

Fonte: Elaborada pelo autor.

Tabela 15 - Resultados obtidos para a métrica de Medida-F – Fator (Base de dados) x Bloco (Percentual de completude da rota), para o modelo PPM-HMM.

Bloco B	***Fator F* – Base de dados**		**Média**
	TodosTrajetos	*TrajetosMaiorQueDois*	
15%	0,31	0,35	0,33
50%	0,36	0,47	0,42
85%	0,39	0,56	0,48
Média alternativas	**0,35**	**0,46**	**0,41**

Fonte: Elaborada pelo autor.

Análise dos Resultados Obtidos

Conforme os resultados obtidos pela análise estatística, é possível refutar a hipótese nula (H2-0) da Questão de Pesquisa 2 (**QP2**) apenas para a métrica de *precisão*, mas não para as métricas de *cobertura* e *medida-F*. Isto é, utilizando o modelo de previsão com base em PPM-HMM, não é possível afirmar que há diferença na média dos valores obtidos para as métricas de *cobertura* e *medida-F*, ao variar as alternativas do *fator* base de dados (com as alternativas *TodosTrajetos* e *TrajetosMaiorQueDois*). Com o resultado desta análise, poderia ser sugerido que o PPM-HMM obteve resultados semelhantes tanto para o cenário de testes (1), como para o cenário de testes (2), para as métricas de *cobertura* e *medida-F*¸o que o tornaria robusto à variação da base de dados utilizada. Porém, ao analisar a influência nos resultados obtidos, percebe-se que a influência do *bloco* foi superior à influência do *fator* nos resultados obtidos.

Fazendo uma comparação entre os resultados obtidos com os modelos PPM-Markov e PPM-HMM, é possível afirmar que, estatisticamente, para o modelo PPM-Markov, o uso da base de dados (com os níveis *TodosTrajetos* e *TrajetosMaiorQueDois*) influenciou mais no resultado (com índices de influência de 98% para todas as variáveis resposta) do que com o uso do modelo PPM-HMM. Essa afirmação é possível mediante a análise de que o resultado obtido para a *precisão* foi mais influenciado pelo *bloco B* do que o *fator F*, ao considerar o modelo de previsão PPM-HMM.

5.3.2.3. Testes com o Modelo de Previsão PPM

Esta seção apresenta os experimentos agrupados pelo modelo de previsão PPM puro, ou seja, sem combinação do PPM com qualquer outra técnica. As Tabelas 16, 17 e 18 apresentam, respectivamente, os resultados obtidos para as métricas de *precisão*, *cobertura* e *medida-F*, após a realização dos experimentos. A primeira coluna das tabelas

é destinada aos níveis do *bloco B* (exceto as duas últimas linhas), enquanto a segunda coluna apresenta os resultados das médias obtidos para a métrica em análise, referente aos dois níveis do *fator F*. A última coluna apresenta as médias obtidas por bloco, e a última linha apresenta as médias de cada uma das duas alternativas do *fator F*.

Tabela 16 - Resultados obtidos para a métrica de *Precisão* – Fator (Base de dados) x Bloco (Percentual de completude da rota), para o modelo PPM.

Bloco B	***Fator F* – Base de dados**		**Média**
	TodosTrajetos	*TrajetosMaiorQueDois*	
15%	0.47	0.62	0.55
50%	0.54	0.67	0.61
85%	0.64	0.74	0.69
Média	**0.55**	**0.68**	**0.61**

Fonte: Elaborada pelo autor.

Tabela 17 - Resultados obtidos para a métrica de Cobertura – Fator (Base de dados) x Bloco (Percentual de completude da rota), para o modelo PPM.

Bloco B	***Fator F* – Base de dados**		**Média**
	TodosTrajetos	*TrajetosMaiorQueDois*	
15%	0.36	0.63	0.50
50%	0.48	0.73	0.61
85%	0.63	0.79	0.71
Média	**0.49**	**0.72**	**0.60**

Fonte: Elaborada pelo autor.

Tabela 18 - Resultados obtidos para a métrica de Medida-F – Fator (Base de dados) x Bloco (Percentual de completude da rota), para o modelo PPM.

Bloco B	***Fator F* – Base de dados**		**Média**
	TodosTrajetos	*TrajetosMaiorQueDois*	
15%	0.41	0.63	0.52
50%	0.51	0.695	0.60
85%	0.64	0.76	0.70
Média	**0.52**	**0.70**	**0.61**

Fonte: Elaborada pelo autor.

Análise dos Resultados Estatísticos

Conforme os resultados obtidos para as variáveis resposta *precisão, cobertura* e *medida-F*, é possível refutar a hipótese nula (H2-0) da questão de pesquisa 2 (**QP2**), sob confiança de 95%, e afirmar que não há evidência estatística de que as médias dos valores obtidos para as bases de dados *TodosTrajetos* e *TrajetosMaiorQueDois* são iguais, também para o modelo de previsão baseado apenas no PPM. Quanto à influência nos

resultados obtidos, os testes mostram que o *fator F* é mais influente do que o *bloco B*, no entanto, os valores de influência do fator não são tão superiores do que os do bloco, principalmente para as métricas *precisão* e *medida-F*. Isto demonstra que, embora a confiança de que o *fator F* seja mesmo mais influente nos resultados do que o *bloco B* obtidos para as métricas estatísticas avaliadas, o *bloco B* também tem considerável influência na obtenção dos resultados.

5.3.2.4. Análise Subjetiva dos Testes para a Questão de Pesquisa 2

O objeto de análise da questão de pesquisa 2 (**QP2**) é verificar a influência do fator *base de dados* na obtenção dos resultados de previsão de trajetórias. Isto é, buscou-se investigar qual a influência de uma base de dados preenchida, majoritariamente, com rotas percorridas uma única vez (base *TodosTrajetos*) comparada com uma base de dados filtrada com rotas que foram realizadas pelo menos duas vezes (base *TrajetosMaiorQueDois*). Para isso, os dados foram blocados conforme o *percentual de completude da rota a ser testada*. Após isso, os testes foram realizados separadamente conforme os três modelos de previsão com base no PPM (PPM-Markov, PPM-HMM e PPM). As métricas avaliadas foram *precisão, cobertura* e *medida-F*.

Dentre os modelos de previsão avaliados, o modelo PPM-Markov manteve os resultados convergentes para as três métricas, tendo o fator influenciado em mais de 90% nos resultados estatísticos. Esta informação demonstra uma confiança alta de que a base de dados influencia nos resultados obtidos mais do que o bloco, quando se trata do uso de PPM-Markov, isto é, o preditor PPM-Markov tende a melhorar a previsão da rota e do destino em uma base de dados que possui um maior padrão de deslocamentos. Além disso, os testes mostraram que também foi possível refutar a hipótese nula H2-0, de que não há diferença nas bases *TodosTrajetos* e *TrajetosMaiorQueDois* para previsão, sob confiança de 95%. Se a confiança fosse analisada sob 99%, o resultado seria o mesmo, ou seja, ainda seria possível refutar a hipótese nula H2-0 para PPM-Markov.

Quanto ao modelo PPM-HMM, foi possível refutar a hipótese nula H2-0 apenas para a métrica *precisão*, mas não para as métricas de *cobertura* e de *medida-F*. No entanto, para a variável resposta de *precisão*, o bloco foi mais influente do que o fator, obtendo, respectivamente, valores de aproximadamente 61% e 39% de influência. Esta análise demonstra que não foi o fator *base de dados* que influenciou no resultado, mas o

bloco *percentual de completude da rota a ser testada*. Com relação às métricas de *cobertura* e *medida-F*, embora não tenha sido possível refutar a hipótese nula, o *bloco* também apresentou maior influência nos resultados do que o *fator*. Com isso, comparando-se a confiança da influência nos resultados dos modelos de previsão PPM-Markov *versus* PPM-HMM, há evidências estatísticas de que aquele seja mais influente do que este.

Com o modelo de previsão PPM, também foi possível refutar a hipótese nula H2-0, assim como ocorreu no modelo PPM-Markov. No entanto, a influência do fator e do bloco ficou muito próxima para as métricas de *precisão* e de *medida-F*. Ou seja, há evidências estatísticas de que, embora o fator tenha influenciado mais os resultados, eles não foram tão mais influentes do que o bloco e, portanto, não houve tanta diferença nos resultados quando variou-se da alternativa do fator *TodosTrajetos* para *TrajetosMaiorQueDois*. Já a *cobertura* apresentou uma diferença maior, com 88% de influência do fator.

Fazendo-se uma comparação entre os modelos de previsão, para a base de dados *TodosTrajetos*, a maior *precisão* obtida foi com o modelo PPM-Markov (0,67, com 85% da rota percorrida) seguido por PPM (0,64, com 50% da rota percorrida). Quanto à *cobertura*, o maior resultado alcançado foi para o modelo PPM, com um valor de 0,63, valor muito superior às *coberturas* de PPM-Markov (0,44) e de PPM-HMM (0,39). Uma investigação interessante seria tentar utilizar o modelo PPM para *cobertura*, isto é, para a recuperação das trajetórias candidatas à previsão, e utilizar o modelo PPM-Markov para a previsão.

Quando à base *TrajetosMaiorQueDois*, o modelo de previsão PPM-Markov foi superior em todas as métricas analisadas, quando comparado com os outros modelos. Para *precisão*, PPM-Markov obteve índices que variaram de 0,77 a 0,82, conforme aumenta o percentual de completude da rota a ser testada. Ou seja, já no início de um trajeto (15% da rota percorrida), o percentual de precisão chega a ser de 77%, com uma *cobertura* de 0,86. Para 85% da rota percorrida, a *cobertura* alcançou o valor de 0,91 e a *precisão* de 0,82, resultando em uma *medida-F* de 0,87. O modelo de previsão PPM proporcionou resultados semelhantes para as métricas avaliadas. Como o PPM-Markov apresentou evidência estatística de ser mais influente do que os demais modelos na obtenção dos resultados, para uma aplicação prática, este modelo de previsão aparenta ser o mais indicado.

5.3.3. Comparação com a Literatura

Nesta subseção, os resultados dos algoritmos de previsão propostos e desenvolvidos nesta pesquisa (que são baseados em PPM) são comparados com resultados de outras abordagens de previsão disponíveis na literatura. Para comparação, foi utilizada a métrica de *precisão*, conforme ilustrada na Figura 10, uma vez que a precisão é uma métrica presente na maioria dos artigos. Herder et al. (2014) testaram o modelo desenvolvido por eles com a mesma base de dados usada para avaliação dos algoritmos propostos nesta pesquisa (porém, eles também utilizaram a base de dados do projeto *GeoLife*). Para a previsão do próximo destino, eles obtiveram uma precisão de 0,626, utilizando um preditor baseado em Markov. O modelo proposto por Krumm (2008), que prevê até *n* segmentos adiante (da localização corrente do usuário), obteve uma precisão de 0,5 quando *n = 10*. O modelo de previsão apresentado em Simmons et al. (2006) alcança uma taxa de precisão que varia entre 0,7 e 0,8, quando há a ocorrência de um entroncamento na localização corrente do usuário. Usando um procedimento de teste similar ao implementado por este trabalho (segmentar em porcentagens as rotas a serem testadas – por exemplo, em 15%, 50% e 85%), o preditor desenvolvido por Froehlich e Krumm (2008) obteve uma precisão máxima de 0,43, quando é informada 85% da rota a ser testada. No trabalho de Lee et al. (2016), foi desenvolvido um preditor baseado no algoritmo GSTP, e foi alcançada uma precisão de 0,74. Ying, Lee e Weng (2011) reportaram uma precisão de 0,68 e medida-F de 0,78 no preditor desenvolvido por eles (utiliza informação semântica e prevê apenas o próximo destino). O modelo de previsão desenvolvido por Lung, Chung e Daí (2014), que também utiliza informação semântica, alcançou uma precisão de 0,683. Com relação aos resultados dos algoritmos de previsão propostos neste trabalho, PPM-Markov proporcionou precisão de 0,66 e de PPM 0,64 para a base de dados *TodosTrajetos*, quando é informada 85% da rota a ser testada. Para a base *TrajetosMaiorQueDois*, a maior precisão obtida para o algoritmo PPM-Markov foi de 0,82 e para o algoritmo PPM foi de 0,74.

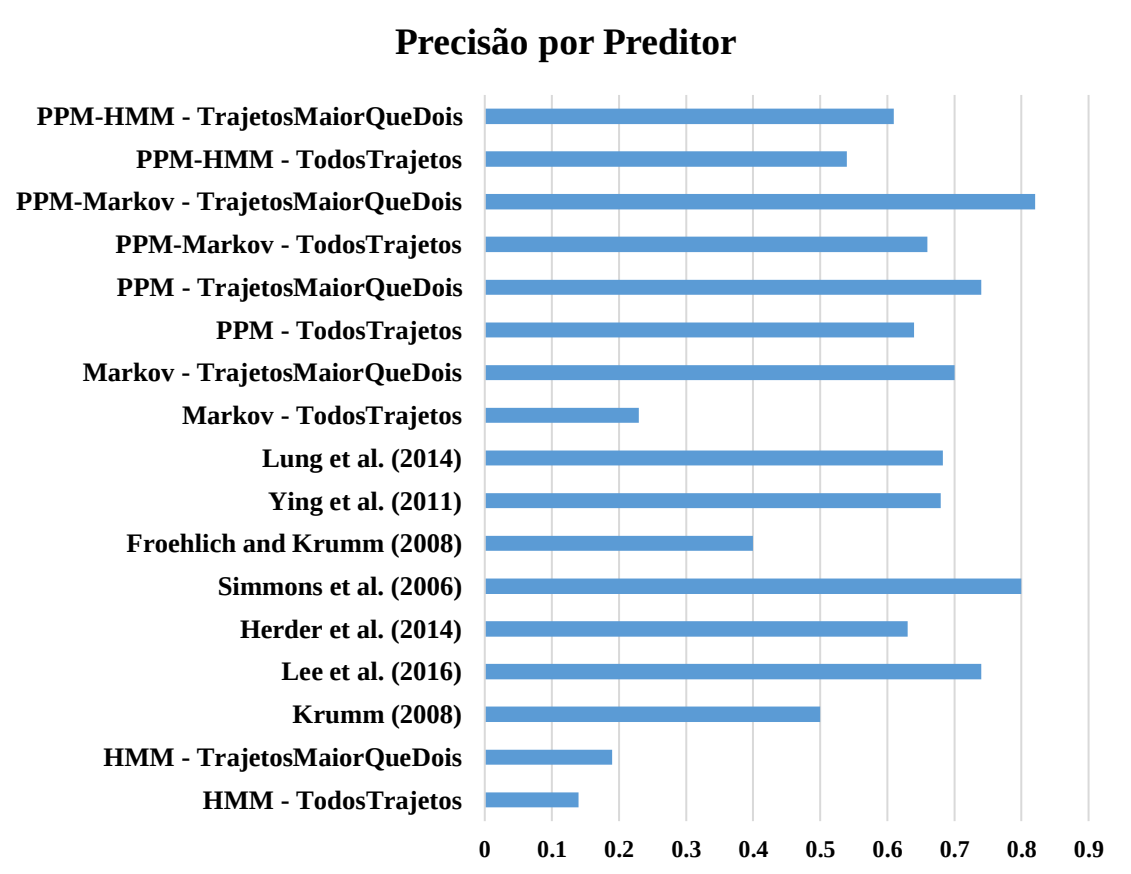
5.4. Conclusão do Capítulo

Neste capítulo, foram apresentadas as análises estatísticas referentes às questões de pesquisa **QP1** e **QP2**. Com relação à **QP1**, há evidência estatística de que os modelos avaliados são diferentes e influenciam na qualidade do resultado da previsão de destino,

com relação às métricas *previsão*, *cobertura* e *medida-F*. Dentre estes, os modelos de previsão baseados em PPM (PPM-Markov, PPM-HMM e PPM) apresentaram resultados competitivos àqueles obtidos pelos modelos de previsão Markov e HMM, desenvolvidos de maneira preliminar, com relação às variáveis resposta analisadas. Quanto à **QP2**, para os modelos de previsão PPM-Markov e PPM, também foi possível refutar a hipótese nula de que o *uso da base de dados* não influencia na previsão de trajetórias, sob confiança de 95%.

No capítulo a seguir, são apresentadas as conclusões obtidas com o desenvolvimento deste trabalho, em que são ressaltadas as principais contribuições e afirmações baseadas em experimentos. Além disso, são mencionadas oportunidades futuras de pesquisa que se alinhem com este trabalho.

Figura 10 - Comparação dos resultados preliminares dos modelos de previsão propostos com resultados dos modelos descritos na literatura.



Fonte: Elaborada pelo autor.

Capítulo 6

Conclusão e Sugestão para Trabalhos Futuros

As conclusões obtidas com a pesquisa de doutorado relatada nesta tese são apresentadas neste capítulo. É importante notar que as conclusões obtidas se restringem aos dados utilizados para experimento deste trabalho (os dados MSR GPS Privacy Dataset, oriundos de um projeto realizado pela Microsoft), não podendo ser estendidas para qualquer dado referente a deslocamento. Ainda neste capítulo, são apresentadas sugestões para trabalhos futuros que poderão ser realizados nesta área de pesquisa.

6.1. Conclusão

O foco deste trabalho de doutorado foi desenvolver um modelo de previsão de trajetória capaz de prever rotas e destinos que um usuário deve alcançar, mesmo sendo este destino um lugar que ainda não tenha sido visitado, com o auxílio de informações contextuais e semânticas. Além disso, procurou-se desenvolver preditores automatizados para a realização das suas funcionalidades, aumentando, assim, sua utilidade para deslocamentos diários e rotineiros.

Com o objetivo definido, o sistema *Predroute* foi desenvolvido com a finalidade de identificar rotas automaticamente a partir de coordenadas geo-temporais, descobrir os tipos de lugares que o usuário visita, além de realizar previsão de destino e rotas. Foram criadas duas questões de pesquisa para avaliação dos modelos propostos neste trabalho (ver Seção 1.2), além de implementação e comparação dos modelos de previsão propostos

com modelos de previsão já consolidados na literatura, como aqueles que são baseados em Cadeias de Markov e em Modelos Ocultos de Markov (HMM). As conclusões referentes ao trabalho de doutorado desenvolvido são descritas a seguir:

- A principal contribuição deste trabalho foi apresentar uma família de preditores baseados em PPM, a saber: (1) um preditor com PPM puro; (2) outro combinando PPM-Markov; e um (3) terceiro que combina PPM-HMM. Dentre estes modelos propostos, o que combinou PPM e Markov, estatisticamente, obteve mais influência nas obtenções das variáveis resposta comparados aos modelos baseados em PPM e HMM, e PPM puro, para ambas as bases de dados: *TodosTrajetos* e *TrajetosMaiorQueDois*. Estes dois últimos modelos (PPM-HMM e PPM puro) dividiram parte da influência no resultado com o bloco referente ao *percentual de completude da rota a ser testada*;
- Em virtude do modelo de previsão PPM-Markov ter obtido resultados encorajadores e ser mais confiável no quesito de influência nas métricas, ele aparenta ser o preditor mais robusto dentre os que foram avaliados. Além disso, os melhores resultados obtidos (para as métricas avaliadas) foram quando o percentual de completude da rota esteve em 85%. Deve-se destacar que, quando 85% da rota foi percorrida, geralmente, o usuário está próximo de alcançar seu destino. Mesmo neste momento (85% da rota percorrida) a previsão de destino é importante, pois pode alertar que o destino está congestionado (por exemplo, por uma manifestação ou por um evento – concurso público, concertos, entre outros);
- Embora o foco desta pesquisa tenha sido a proposta de modelos de previsão baseados em PPM, também foram implementados módulos para identificação de paradas (lugares visitados), criação de rotas e enriquecimento semântico. Todos estes módulos poderiam ser substituídos por outros (com pesquisas mais aprofundadas nas respectivas áreas) sem impacto nos preditores propostos;
- O mecanismo de *rota reduzida* – também uma contribuição deste trabalho – tem uma proposta importante: identificar a realização de uma rota pela primeira vez. Com este mecanismo, os resultados para as métricas estatísticas da família de preditores baseados no PPM (já que foram implementados somente nestes preditores) foram superiores aos modelos de Markov e HMM, ainda que mediante a uma implementação preliminar destes dois últimos modelos. No entanto, uma

pesquisa mais aprofundada deve ser conduzida, pois, comprovando sua eficácia, essa estratégia pode ser aplicada em áreas diferentes da de previsão de destino;

- Há considerável contribuição na área de previsão de destino. Porém, boa parte dos trabalhos da área desconsidera enriquecimento semântico (o papel que um lugar representa), o que apresenta uma oportunidade de pesquisa e um campo vasto ainda a ser explorado. Nos últimos anos, no entanto, pesquisadores têm explorado mais o uso de semântica para anotar trajetórias;
- Por meio dos resultados experimentais, outra contribuição importante do trabalho foi a possibilidade de verificar, estatisticamente, que o comportamento de deslocamento dos usuários possui considerável influência nas métricas de *precisão*, *cobertura* e *medida-F*.

6.2. Sugestão para Trabalhos Futuros

Esta pesquisa propôs a investigação de questões relacionadas à área de pesquisa de previsão de trajetórias, principalmente, sugerindo modelos que ajustem sua previsão conforme o progresso de deslocamento do usuário (os que são baseados em PPM) e o mecanismo de *rota reduzida*, que identifica que uma rota em curso nunca foi realizada anteriormente. Como os modelos de previsão baseados em PPM apresentam evidência estatística de serem robustos, principalmente combinados com outros métodos, os próximos passos incluem o aperfeiçoamento deste modelo, dos testes e de investigação de novas informações contextuais. A seguir, são sugeridos alguns trabalhos futuros, que podem ser originados a partir desta pesquisa:

- Os modelos de previsão baseados em PPM apresentaram resultados compatíveis e superiores com os da literatura, como, por exemplo, os modelos baseados em Markov e HMM. No entanto, ainda há um espaço a ser explorado principalmente com relação à métrica de *cobertura*, quando a base é composta majoritariamente por trajetos realizados apenas uma vez;
- Os modelos de previsão Markov e HMM implementados, nesta pesquisa, possuem a capacidade de prever apenas o próximo destino, ao invés de toda a rota restante. Uma investigação futura importante, principalmente para o HMM, é criar uma modelagem capaz de contemplar, além das informações contextuais de uma trajetória, os segmentos percorridos em uma rota, para melhorar a previsão;

- A combinação de técnicas aparentou ter sido apropriada para obter melhores valores para as métricas estatísticas. Portanto, para previsão de trajetórias, pretende-se investigar métodos como, por exemplo, *Active Learning*;
- Toda vez que uma rota é comprimida com uma árvore de símbolos PPM, um número referente à razão de compressão (RC) é gerado. Uma atividade futura consiste em investigar se há alguma correlação entre as RCs geradas para verificar se houve confusão na tentativa de acertar o trajeto previsto ou se o algoritmo possui razoável certeza. Isto tende a aumentar a *precisão*, mas pode diminuir a *cobertura*, principalmente em bases de dados com muitos deslocamentos realizados apenas uma vez;
- Planeja-se investigar o impacto de novas informações contextuais nas previsões de trajetórias. Por exemplo, informações que se referem às características do usuário, como, por exemplo, gênero, tipos de lazer de preferência, idade, entre outros, poderão ser utilizadas para aperfeiçoamento da previsão;
- O mecanismo de previsão de lugares nunca visitados se restringe à previsão de um lugar com o mesmo papel de lugar do destino inicialmente previsto. Sugere-se, como trabalho futuro, ampliar este mecanismo para prever lugares novos mesmo que o papel de lugar seja diferente do papel de lugar referente ao destino inicialmente previsto;
- Pretende-se, ainda, aprofundar os testes com o mecanismo de rota reduzida, que pode ajudar, ainda no início do deslocamento, no discernimento se uma rota que está sendo realizada é totalmente nova. De posse dessa informação, o modelo poderia fazer uso de outras áreas de conhecimento, como sugestão de POIs (YU e CHEN, 2016) e *descoberta de informação valiosa ao acaso* (*Serendipity*).

REFERÊNCIAS BIBLIOGRÁFICAS

ABE, N.; WARMUTH, M. K. On the Computational Complexity of Approximating Distributions by Probabilistic Automata. **Machine Learning**, Julho 1992. 205-260.

ANDREZZA, I. L. P.; BORGES, E. V. C. D. L.; BATISTA, L. V. Heart arrhythmia classification using the prediction by partial matching algorithm. **International Journal of Computer Applications in Technology**, 2015. 285-291.

AREL, I.; ROSE, D. C.; KARNOWSKI, T. P. Deep Machine Learning - A New Frontier in Artificial Learning Research. **IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE**, p. 13-18, 2010.

BIRANT, D.; KUT, A. ST-DBSCAN: An algorithm for clustering spatial–temporal data. **Data & Knowledge Engineering**, v. 60, n. 1, p. 208-221, 2007.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

BREJOVÁ, B.; BROWN, D. G.; VINAř, T. **The Most Probable Labeling Problem in HMMs and Its Application to Bioinformatics**. International Workshop on Algorithms in Bioinformatics. [S.l.]: Springer. 2004. p. 426-437.

BRILHANTE, I. R. et al. On planning sightseeing tours with TripBuilder. **Inf. Process. Manage**, 2015. 1-15.

BRUSH, A.; KRUMM, J.; SCOTT, J. Exploring End User Preferences for Location Obfuscation, Location-Based Services, and the Value of Location. **UbiComp'10**, 2010. 95-104.

BURBEY, I.; MARTIN, T. L. Predicting Future Locations Using Prediction-by-Partial-Match. **Workshop on Mobile Entity Localization and Tracking in GPS-less Environments**, 2008. 1-6.

CHEN, P. P.-S. The Entity-Relationship Model - Toward a Unified View of Data. **ACM Transactions on Database Systems**, v. 1, n. 1, p. 9-36, 1976.

CONSÓRCIO DE INSTITUIÇÕES (CNR; UNIVE; UPRC; UNB; UFC; UFPE; UFSC; PUC-RIO). SEmantic Enrichment of trajectory Knowledge discovery - SEEK Project, 2012. Disponivel em: <http://www.seek-project.eu/>. Acesso em: Fevereiro 2016.

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Machine Learning**, p. 273-297, 1995.

COWLES, M.; DAVIS, C. On the origins of the.05 level of statistical significance. **American Psychologist**, Maio 1982. 553-558.

EAGLE, N.; PENTLAND, A. (. Reality mining: sensing complex social systems. **Journal Personal and Ubiquitous Computing**, v. 10, n. 4, p. 255-268, 2006.

ESTER, M. et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. **KDD Proceedings**, 1996. 226-231.

FEI, X.; LUB, C.-C.; LIUC., K. A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. **Transportation Research Part C: Emerging Technologies**, v. 19, n. 6, p. 1306-1318, 2011.

FIGUEIREDO, F. et al. TribeFlow: Mining & Predicting User Trajectories. **Proceedings of the 25th International Conference on World Wide Web**, Montreal, Abril 2016. 695-706.

FROEHLICH, J.; KRUMM, J. Route Prediction from Trip Observations. **Society of Automotive Engineers (SAE)**, Abril 2008.

GEORGESCU, L.; ZEITLER, D.; STANDRIDGE, C. R. Intelligent transportation system real time traffic speed prediction with minimal data. **Journal of Industrial Engineering and Management**, v. 5, n. 2, p. 431-441, 2012.

GILLMAN, D.; SIPSER, M. **Inference and minimization of hidden Markov chains**. Proceedings of the seventh annual conference on Computational learning theory. New Brunswick: ACM. 1994. p. 147-158.

HAN, J.; KAMBER, M.; TUNG, A. Spatial clustering methods in data mining: a survey. **Geographic Data Mining and Knowledge Discovery**, Inc. Bristol, PA, n. 1, Janeiro 2001. ISSN 0415233690.

HERDER, E.; SIEHNDEL, P.; KAWASE, R. **Predicting User Locations and Trajectories**. 22nd International Conference, UMAP 2014. Aalborg, Denmark: Springer. 2014. p. 86-97.

HONORIO, T. C. S.; BATISTA, L. V.; DUARTE, R. C. M. **Texture Classification Using Prediction by Partial Matching Models**. Workshop de Visão Computacional. São Paulo: Anais do V Workshop de Visão Computacional. 2009. p. 21-26.

HUANG, C.-M.; YING, J. J.-C.; TSENG, V. S. **Mining Users' Behaviors and Environments for Semantic Place Prediction**. Mobile Data Challenge by Nokia Workshop. Newcastle: . 2012. p. 6.

HUTTENLOCHER, D. P.; RUCKLIDGE, W. J. **A Multi-Resolution Technique for Comparing Images using the Hausdorff Distance**. Cornell University. Ithaca, p. 23. 1992.

JURAFSKY, D.; H. MARTIN, J. **Speech and Language Processing**. 2a. ed. : Prentice Hall, 2008.

JURISTO, N.; MORENO, A. M. **Basics of Software Engineering Experimentation**. 1a. ed. : Springer, 2010.

KANUNGO, T. et al. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. **IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE**, v. 24, n. 7, p. 881-892, 2002.

KIMBAL, R.; ROSS, M. **The Data Warehouse Toolkit**. [S.l.]: John Wiley, 2002.

KISILEVICH, S. et al. Spatio-Temporal Clustering: a Survey. **Data mining and Knowledge Discovery Handbook**, New York, p. 855-874, 2010.

KRUMM, J. A Markov Model for Driver Turn Prediction. **Society of Automotive Engineers (SAE)**, Abril 2008.

LAURILA, J. K. et al. **The Mobile Data Challenge:** Big Data for Mobile Computing Research. Proc. Mobile Data Challenge Workshop (MDC) in conjunction with Pervasive. Newcastle: . 2012.

LEE, S. et al. Next Place Prediction Based on Spatiotemporal Pattern Mining of Mobile Device Logs. **Sensors**, Janeiro 2016. 1-19.

LEI, P.; LI, S.; PENG, W. QS-STT: QuadSection clustering and spatial-temporal trajectory model for location prediction. **Distributed and Parallel Databases**, 2013. 231-258.

LIU, X.; LIU, Y.; LI, X. Exploring the context of locations for personalized location recommendations. **Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence**, New York, Julho 2016. 1188-1194.

LUNG, H.-Y.; CHUNG, C.-H.; DAI, B.-R. Predicting Locations of Mobile Users Based on Behavior Semantic Mining. **Trends and Applications in Knowledge Discovery and Data Mining**, v. 8643, p. 168-180, 2014.

MOFFAT, A. Implementing the PPM Data Compression Scheme. **IEEE TRANSACTIONS ON COMMUNICATIONS**, v. 38, n. 11, p. 1917-1921, Novembro 1990.

MONREALE, A. et al. **WhereNext:** a Location Predictor on Trajectory Pattern Mining. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD). Paris: ACM. 2009. p. 637-646.

MONTGOMERY, D. C.; RUNGER, G. C. **Applied Statistics and Probability for Engineers**. 5a. ed. : John Wiley & Sons, 2011.

MORZY, M. Prediction of moving object location based on frequent trajectories. **ISCIS**, 2006. 583-592.

NADEMBEGA , ; TALEB, T.; HAFID, A. A Destination Prediction Model based on historical data, contextual knowledge and spatial conceptual maps. **2012 IEEE International Conference on Communications (ICC)**, Junho 2012. 1416-1420.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in Neurorobotics**, 2013. Disponivel em:

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/>. Acesso em: Fevereiro 2016.

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM. **Travel Demand Forecasting: Parameters and Techniques**. Washington, DC. 2012.

NORRIS, J. R. **Markov Chains (Cambridge Series in Statistical and Probabilistic Mathematics)**. 1a. ed. : Cambridge University Press, 1998.

PARENT, C. et al. Semantic trajectories modeling and analysis. **ACM Computing Surveys**, v. 45, n. 4, p. 1-32, 2013.

PAVELEC, D. et al. **Compression and stylometry for author identification**. International Joint Conference on Neural Networks (IJCNN). Atlanta: IEEE Press Piscataway. 2009. p. 2445-2450.

PENG, C.-Y. J.; LEE, K. L.; INGERSOLL, G. M. An Introduction to Logistic Regression Analysis and Reporting. **The Journal of Education Research**, p. 3-14, 2002.

PERERA, C. et al. Context Aware Computing for The Internet of Things: A Survey. **IEEE Communications Surveys & Tutorials Journal**, 2013.

PLATT, J. C. **Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines**. MIT Press. , p. 21. 1998. (MSR-TR-98-14).

QUDDUS, M. A.; NOLAND, R. B. A high accuracy fuzzy logic based map matching algorithm for road transport. **Journal of Intelligent Transportation Systems**, v. 10, n. 3, p. 103-115, 2006.

QUILAN, J. R. **C4.5:** Programs for Machine Learning. San Francisco: Morgan Kaufmann, 1993.

RAZALI, N. M. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. **Journal of Statistical Modeling and Analytics**, Janeiro 2011. 21-33.

ROCHA, C. L. et al. **TPRED:** a Spatio-Temporal Location Predictor Framework. Proceedings of the 20th International Database Engineering & Applications (IDEAS). Montreal: ACM. 2016. p. 34-42.

RON, D.; SINGER, Y.; TISHBY, N. The power of amnesia: Learning probabilistic automata with variable memory length. **Machine Learning**, 1996. 117-149.

SALOMON, D. **Data Compression:** The Complete Reference. 3a. ed. : Springer, 2004.

SILVA, T. L. C. D.; MACÊDO, J. A. F. D.; CASANOVA, M. A. Discovering Frequent Mobility Patterns on Moving Object Data. **Proceedings of the Third ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems**, Novembro 2014. 60-67.

SIMMONS, R. et al. Learning to Predict Driver Route and Destination Intent. **Intelligent Transportation Systems Conference**, 2006. 127-132.

SPACCAPIETRA, S. et al. A conceptual view on trajectories. **Data & Knowledge Engineering**, v. 65, n. 1, p. 126-146, 2008.

STAMP, M. **A Revealing Introduction to Hidden Markov Models**. San Jose State University. San Jose, p. 21. 2015. (-).

SUSSMAN, J. S. **Perspectives on Intelligent Transportation Systems (ITS)**. [S.l.]: Springer, 2005.

TANAKA, K. et al. A Destination Prediction Method Using Driving Contexts and Trajectory for Car Navigation Systems. **Proceedings of the 2009 ACM symposium on Applied Computing - SAC '09**, 2009. 190-195.

TIWIRI, V. S.; ARYA, A.; CHATUVERDI, S. Route prediction using trip observations and map matching. **IEEE 3rd International Advance Computing Conference (IACC)**, Fevereiro 2013. 583-587.

TORK, H. F. **Spatio-temporal clustering methods classification**. Doctoral Symposium on Informatics Engineering. Porto, Portugal: . 2012. p. 1-12.

TRASARTI, R. et al. MyWay: Location prediction via mobility profilin. **Information Systems**, 06 Novembro 2015. 1-18.

VAN RIJSBERGEN, C. J. **Information Retrieval**. 2ª. ed. Massachusetts: Butterworth-Heinemann Newton, 1979.

VARRIALE, R.; MA, S.; WOLFSON, O. A Volunteered Travelers Information System. **Proceedings of the Sixth ACM SIGSPATIAL International Workshop on Computational Transportation Science**, 2013.

VIEIRA, V.; TEDESCO, P. A.; SALGADO, A. C. Designing context-sensitive systems: An integrated approach. **Expert Systems with Applications**, v. 38, n. 2, p. 1119-1139, 2011.

VLAHOGIANNI, E. I.; KARLAFTIS, M. G.; GOLIAS., J. C. Short-time travel forecasting: Where we are and where we're going. **Transportation Research Part C: Emerging Technologies**, v. 43, n. 1, p. 3-19, 2014.

WINTER, S. et al. Towards a Computational Transportation Science. **Journal of Spatial Information Science (JOSIS)**, n. 2, p. 119-126, 2011.

WOLFSON, O.; SISTLA, A. P.; XU, B. The TranQuyl language for data management in intelligent transportation. **Transportation Research Part C: Emerging Technologies**, v. 23, p. 3-13, 19 Agosto 2012.

XUE, A. Y. et al. Destination Prediction by Sub-Trajectory Synthesis and Privacy Protection Against Such Prediction. **IEEE 29th International Conference on Data Engineering (ICDE)**, 2013. 254-265.

YING, J. J.-C.; LEE, W.-C.; WENG, T.-C. Semantic Trajectory Mining for Location Prediction. **Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems**, 2011. 34-43.

YING, J.-C. et al. Semantic trajectory-based high utility item recommendation system. **Expert Systems with Applications**, v. 41, n. 10, p. 4762-4775, 2014.

YU, ; CHEN, X. **A Survey of Point-of-Interest Recommendation in Location-Based Social Networks**. 2015 AAAI Workshop. Austin: AAAI. 2016. p. 53-60.

ZHANG, J. et al. Data-Driven Intelligent Transportation Systems: A Survey. **IEEE Transactions on Intelligent Transportation Systems**, v. 12, n. 4, p. 1624-1629, 2011.

ZHENG, Y. et al. Mining interesting locations and travel sequences from GPS trajectories. **Proceedings of International conference on World Wild Web (WWW 2009)**, 2009. 791-800.

ZHU, Y. et al. Feature engineering for semantic place prediction. **Pervasive and Mobile Computing**, Amsterdam, Dezembro 2013. 772-783.

APÊNDICE A – Detalhamento dos Resultados Obtidos com os Experimentos para a Questão de Pesquisa 1

Neste apêndice, serão detalhados os resultados obtidos com os experimentos realizados para a questão de pesquisa ***QP1***, apresentada na Seção 1.2, que trata do seguinte problema.

QP1 *– Existe diferença no resultado do uso dos modelos de previsão de rotas e destino (implementados neste trabalho), incluindo previsão de lugares nunca visitados, com relação às métricas estatísticas (precisão, cobertura e medida-F) utilizadas para avaliação?*

H1-0: Não há influência entre os resultados obtidos, com relação ao uso dos modelos de previsão implementados, referentes às métricas estatísticas utilizadas para avaliação.

H1-1: Há influência entre os resultados obtidos, com relação ao uso dos modelos de previsão implementados, referentes às métricas estatísticas utilizadas para avaliação.

Conforme já exposto na Seção 5.3.1, foi adotado o *desgin fatorial* (com um único fator) *com blocagem*, em que o *fator F* se refere aos *modelos de previsão* (com cinco níveis) e o *bloco B* representa o *percentual de completude das rotas a serem testadas* (com três níveis). Os níveis do *fator F* são *PPM-Markov*, *PPM-HMM*, *PPM*, *Markov* e *HMM*, enquanto os níveis do *bloco B* são 15%, 50% e 85%. Os testes foram realizados separadamente para cada um dos dois cenários de teste, resultando, portanto, em duas subseções neste Apêndice: uma para contemplar o cenário (1) *TodosTrajetos*; e outra para contemplar o cenário (2) *TrajetosMaiorQueDois*. As próximas duas subseções apresentam, de forma detalhada, os resultados estatísticos, inclusive com aplicação do teste de *Análise de Variância* (ANOVA) e do teste de normalidade.

A.1. Testes com Cenário de Teste (1) *TodosTrajetos*

Nesta subseção, serão apresentados os testes da **QP1** sob a perspectiva da base de dados do cenário (1) *TodosTrajetos*. As variáveis resposta analisadas pelos testes foram *precisão*, *cobertura* e *medida-F*, portanto, para melhorar a legibilidade do documento, os resultados serão apresentados e explanados separadamente para cada variável. Como os

testes foram repetidos cinco vezes, as tabelas apresentarão as médias dos resultados obtidas após a realização dos cinco testes.

Resultados para a Variável Resposta de Precisão

A Tabela 19 apresenta a média dos resultados obtidos para a métrica de *precisão*, após cinco repetições realizadas dos experimentos. As colunas das tabelas apresentam os cinco níveis do *fator F* (modelos de previsão), enquanto as linhas das tabelas representam os três níveis do *bloco B* (percentual de completude da rota a ser testada).

Tabela 19 - Resultados obtidos para a métrica de *Precisão* – Preditores x Percentual da rota, para o cenário de teste (1) TodosTrajetos.

Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,63	0,44	0,47	0,23	0,14	0,38
50%	0,67	0,50	0,54	0,23	0,14	0,42
85%	0,66	0,54	0,64	0,23	0,14	0,44
Média	**0,65**	**0,49**	**0,55**	**0,23**	**0,14**	**0,41**

Fonte: Elaborada pelo autor.

Uma vez que os resultados das médias obtidos entre os níveis do fator do experimento foram diferentes, o próximo passo foi realizar o teste de hipóteses, com o auxílio da ANOVA. Para isso, os dados apropriados para a análise de variância estão disponíveis na Tabela 20, com a apresentação das médias de *precisão* obtidas, referentes aos níveis do *bloco B* e às alternativas do *fator F*. Além disso, nesta mesma tabela, são apresentados os efeitos dos níveis do bloco e do fator.

Tabela 20 - Adequação dos resultados para aplicação da ANOVA para *Precisão*, referente ao Cenário (1) TodosTrajetos.

Bloco B	***Fator* F – Modelos de Previsão**					**Média bloco**	**Efeito bloco**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM		
15%	0,63	0,44	0,47	0,23	0,14	0,38	-0,03 **(β_1)**
50%	0,67	0,50	0,54	0,23	0,14	0,42	0,00 **(β_2)**
85%	0,66	0,54	0,64	0,23	0,14	0,44	0,03 **(β_3)**
Média alternativas	**0,65 (α_1)**	**0,49 (α_2)**	**0,55 (α_3)**	**0,23 (α_4)**	**0,14 (α_5)**	**0,41**	
Efeito alternativas	**0,24**	**0,08**	**0,14**	**-0,18**	**-0,27**		

Fonte: Elaborada pelo autor.

A Tabela 21 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Até este momento, tem-se os valores da variável resposta de *precisão* (variável *y*), os valores dos efeitos do fator (variável α) e do bloco (variável β) e, a partir deste ponto, estão sendo apresentados os erros (variável *e*). Com todos os cálculos realizados, a próxima etapa foi calcular a análise de variância.

Tabela 21 - Erros experimentais para a Precisão, referentes ao cenário de teste (1) TodosTrajetos.

Bloco B	**Erros - *Fator F* – Modelos de Previsão**				
	PPM-Markov	PPM-HMM	PPM	Markov	HMM
15%	0,01	-0,02	-0,05	0,03	0,03
50%	0,01	0,00	-0,01	0,00	0,00
85%	-0,02	0,02	0,06	-0,03	-0,03

Fonte: Elaborada pelo autor.

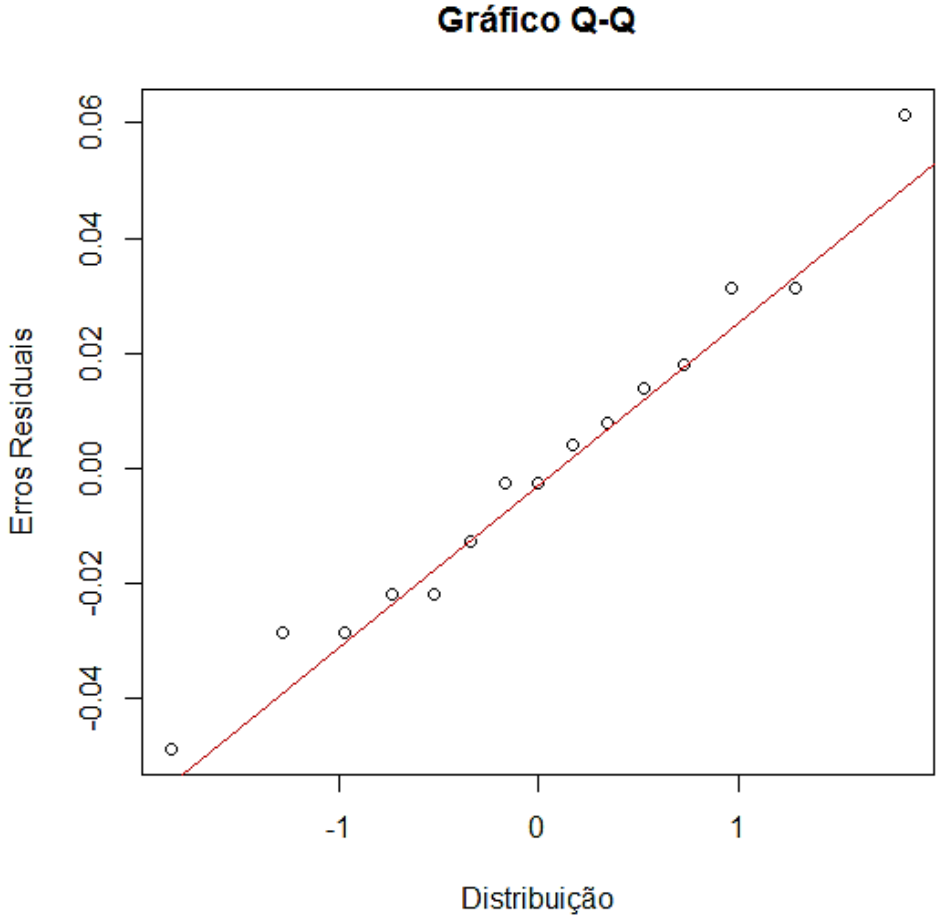
Para o modelo matemático referente à variável resposta *precisão*, tem-se: *SSY = 3,16; SS0 = 2,56; SSB = 0,0091; SSA = 0,5730; SSE = 0,0115; e SST = 0,59*. Com os valores obtidos, obteve-se que o fator influencia, aproximadamente, 96,5%, o bloco influencia 1,5% e os erros também 2%.

Ao utilizar as Equações (2) e (3) obteve-se, respectivamente, que, para esta análise, o valor de MSA é 0,143 e o de MSE é 0,001. Dividindo-se MSA por MSE, tem-se o cálculo de F, igual a 99,826. Como a análise realizada nesta pesquisa requer significância de 95%, a tabela F possui um valor de 19,25. Como o cálculo de F pela análise de variância é maior do que o F da tabela (19,25 < 99,826), é possível refutar, com significância de 95%, a hipótese nula H1-0 da questão de pesquisa ***QP1***. Isto é, pode-se afirmar que, estatisticamente, há diferença nos modelos de previsão avaliados, para prever o destino de uma rota, com respeito à precisão.

Uma vez que a aplicação do teste ANOVA requer que os dados sejam oriundos de uma distribuição normal, deve-se utilizar um teste de normalidade para a verificação dos valores obtidos para a métrica de *precisão*. Ao aplicar o teste de *Shapiro-Wilk* nos erros experimentais, foi possível obter um *valor-p = 0,9515* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

Outra abordagem para verificação da normalidade dos dados é o uso de um gráfico Q-Q, disponível na Figura 11. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Figura 11 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica *precisão*, referente ao cenário de testes (1), da Questão de Pesquisa 1.



Fonte: Elaborada pelo autor.

Resultados para a Variável Resposta de Cobertura

A Tabela 22 apresenta a média dos resultados obtidos para a métrica de *cobertura*, após cinco repetições realizadas dos experimentos. As colunas das tabelas apresentam os cinco níveis do *fator F* (modelos de previsão), enquanto as linhas das tabelas representam os três níveis do *bloco B* (percentual de completude da rota a ser testada).

Tabela 22 – Resultados obtidos para a métrica de cobertura – Preditores x Percentual da rota, para o cenário de teste (1) TodosTrajetos.

Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,32	0,24	0,36	0,30	0,51	0,35
50%	0,38	0,28	0,48	0,30	0,51	0,39
85%	0,44	0,39	0,63	0,30	0,51	0,45
Média	0,38	0,30	0,49	0,30	0,51	**0,40**

Fonte: Elaborada pelo autor.

Uma vez que os resultados das médias obtidos entre os níveis do fator do experimento foram diferentes, o próximo passo foi realizar o teste de hipóteses, com o auxílio da ANOVA. Para isso, os dados apropriados para a análise de variância estão disponíveis na Tabela 23, com a apresentação das médias de *cobertura* obtidas, referentes aos níveis do *bloco B* e às alternativas do *fator F*. Além disso, nesta mesma tabela, são apresentados os efeitos dos níveis do bloco e do fator.

A Tabela 24 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Até este momento, tem-se os valores da variável resposta de *cobertura* (variável *y*), os valores dos efeitos do fator (variável α) e do bloco (variável β) e, a partir deste ponto, estão sendo apresentados os erros (variável *e*). Com todos os cálculos realizados, a próxima etapa foi calcular a análise de variância.

Para o modelo matemático referente à variável resposta *cobertura*, tem-se: *SSY = 2,54; SS0 = 2,36; SSB = 0,03; SSA = 0,12; SSE = 0,026; e SST = 0,18.* Com os valores obtidos, obteve-se que o fator influencia, aproximadamente, 68,2%, o bloco influencia 16,8% e os erros 15%.

Tabela 23 - Adequação dos resultados para aplicação da ANOVA para Cobertura, referente ao Cenário (1) TodosTrajetos.

Bloco B	***Fator F* – Modelos de Previsão**					**Média bloco**	**Efeito bloco**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM		
15%	0,32	0,24	0,36	0,30	0,51	0,35	-0,05 **(β_1)**
50%	0,38	0,28	0,48	0,30	0,51	0,39	0,01 **(β_2)**
85%	0,44	0,39	0,63	0,30	0,51	0,45	0,06 **(β_3)**
Média alternativas	0,38 **(α_1)**	0,30 **(α_2)**	0,49 **(α_3)**	0,30 **(α_4)**	0,51 **(α_5)**	**0,40**	
Efeito alternativas	**-0,02**	**-0,09**	**0,09**	**-0,10**	**0,11**		

Fonte: Elaborada pelo autor.

Tabela 24 - Erros experimentais para a Cobertura, referentes ao cenário de teste (1) TodosTrajetos.

Bloco B	**Erros - *Fator F* – Modelos de Previsão**				
	PPM-Markov	PPM-HMM	PPM	Markov	HMM
15%	-0,01	-0,01	-0,08	0,05	0,05
50%	0,01	-0,02	0,00	0,01	0,01
85%	0,00	0,03	0,08	-0,06	-0,06

Fonte: Elaborada pelo autor.

Ao utilizar as Equações (2) e (3) obteve-se, respectivamente, que, para esta análise, o valor de MSA é 0,0299 e o de MSE é 0,003. Dividindo-se MSA por MSE, tem-se o cálculo de F, igual a 9. Como a análise realizada nesta pesquisa requer significância de 95%, a tabela F possui um valor de 19.25. Como o cálculo de F pela análise de variância é menor do que o F da tabela (19.25 > 9), não é possível refutar, com significância de 95%, a hipótese nula H1-0 da questão de pesquisa ***QP1***, referente à variável resposta de *cobertura*. Isto é, não é possível afirmar que, estatisticamente, há diferença nos modelos de previsão avaliados, para prever o destino de uma rota, com respeito à cobertura.

Uma vez que a aplicação do teste ANOVA requer que os dados sejam oriundos de uma distribuição normal, deve-se utilizar um teste de normalidade para a verificação dos valores obtidos para a métrica de *cobertura*. Ao aplicar o teste de *Shapiro-Wilk* nos erros experimentais, foi possível obter um *valor-p = 0,7008* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

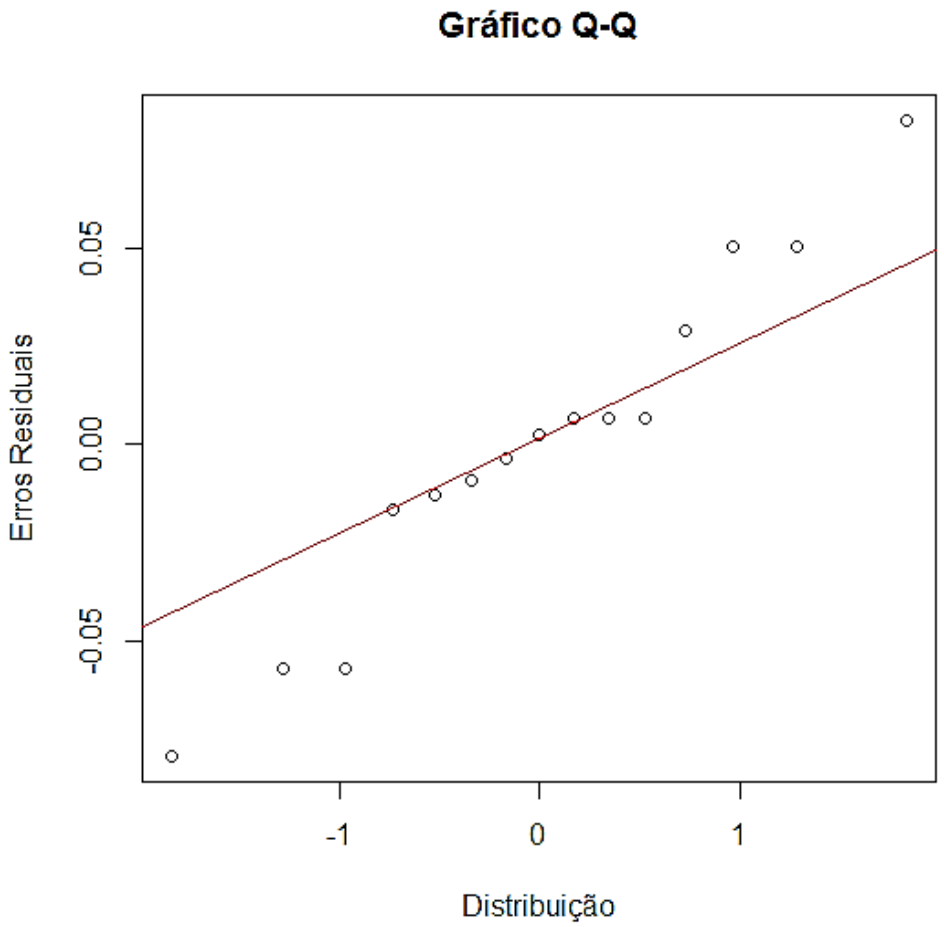
Outra abordagem para verificação da normalidade dos dados é o uso de um gráfico Q-Q, disponível na Figura 12. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Resultados para a Variável Resposta de Medida-F

A Tabela 25 apresenta a média dos resultados obtidos para a métrica de *medida-F*, após cinco repetições realizadas dos experimentos. As colunas das tabelas apresentam

os cinco níveis do *fator F* (modelos de previsão), enquanto as linhas das tabelas representam os três níveis do *bloco B* (percentual de completude da rota a ser testada).

Figura 12 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica cobertura, referente ao cenário de testes (1), da Questão de Pesquisa 1.



Fonte: Elaborada pelo autor.

Tabela 25 - Resultados obtidos para a métrica de medida-F – Preditores x Percentual da rota, para o cenário de teste (1) TodosTrajetos.

Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,42	0,31	0,41	0,23	0,21	0,32
50%	0,49	0,36	0,51	0,23	0,21	0,36
85%	0,53	0,39	0,64	0,23	0,21	0,40
Média	0,48	0,35	0,52	0,23	0,21	**0,36**

Fonte: Elaborada pelo autor.

Uma vez que os resultados das médias obtidos entre os níveis do fator do experimento foram diferentes, o próximo passo foi realizar o teste de hipóteses, com o auxílio da ANOVA. Para isso, os dados apropriados para a análise de variância estão disponíveis na Tabela 26, com a apresentação das médias das *medida-F* obtidas,

referentes aos níveis do *bloco B* e às alternativas do *fator F*. Além disso, nesta mesma tabela, são apresentados os efeitos dos níveis do bloco e do fator.

Tabela 26 - Adequação dos resultados para aplicação da ANOVA para *Medida-F*, referente ao Cenário (1) TodosTrajetos.

Bloco B	***Fator F* – Modelos de Previsão**					**Média bloco**	**Efeito bloco**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM		
15%	0,42	0,31	0,41	0,23	0,21	0,32	-0,04 **(β_1)**
50%	0,49	0,36	0,51	0,23	0,21	0,36	0,00 **(β_2)**
85%	0,53	0,39	0,64	0,23	0,21	0,40	0,04 **(β_3)**
Média alternativas	0,48 **(α_1)**	0,35 **(α_2)**	0,52 **(α_3)**	0,23 **(α_4)**	0,21 **(α_5)**	**0,36**	
Efeito alternativas	**0,12**	**-0,01**	**0,16**	**-0,13**	**-0,15**		

Fonte: Elaborada pelo autor.

A Tabela 27 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Até este momento, tem-se os valores da variável resposta de *medida-F* (variável *y*), os valores dos efeitos do fator (variável α) e do bloco (variável β) e, a partir deste ponto, estão sendo apresentados os erros (variável *e*). Com todos os cálculos realizados, a próxima etapa foi calcular a análise de variância.

Tabela 27 - Erros experimentais para a Medida-F, referentes ao cenário de teste (1) TodosTrajetos.

Bloco B	**Erros - *Fator F* – Modelos de Previsão**				
	PPM-Markov	PPM-HMM	PPM	Markov	HMM
15%	-0,02	0,00	-0,07	0,04	0,04
50%	0,01	0,01	-0,01	0,00	0,00
85%	0,01	0,00	0,08	-0,04	-0,04

Fonte: Elaborada pelo autor.

Para o modelo matemático referente à variável resposta *medida-F*, tem-se: *SSY = 2,2; SS0 = 1,93; SSB = 0,02; SSA = 0,24; SSE = 0,02; e SST = 0,27*. Com os valores obtidos, obteve-se que o fator influencia, aproximadamente, 86,85%, o bloco influencia 6,43% e os erros 6,72%.

Ao utilizar as Equações (1) e (2) obteve-se, respectivamente, que, para esta análise, o valor de MSA é 0,06 e o de MSE é 0,0023. Dividindo-se MSA por MSE, tem-se o cálculo de F, igual a 26. Como a análise realizada nesta pesquisa requer significância

de 95%, o F_c a tabela F possui um valor de 19,25. Como o cálculo de F pela análise de variância é maior do que o F da tabela (19,25 < 26), é possível refutar, com significância de 95%, a hipótese nula H1-0 da questão de pesquisa ***QP1***, referente à variável resposta de *medida-F*. Isto é, é possível afirmar que, estatisticamente, há diferença nos modelos de previsão avaliados, para prever o destino de uma rota, com respeito à medida-F.

Uma vez que a aplicação do teste ANOVA requer que os dados sejam oriundos de uma distribuição normal, deve-se utilizar um teste de normalidade para a verificação dos valores obtidos para a métrica de *medida-F*. Ao aplicar o teste de *Shapiro-Wilk* nos erros experimentais, foi possível obter um *valor-p = 0,5567* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

Outra abordagem para verificação da normalidade dos dados é o uso de um gráfico Q-Q, disponível na Figura 13. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

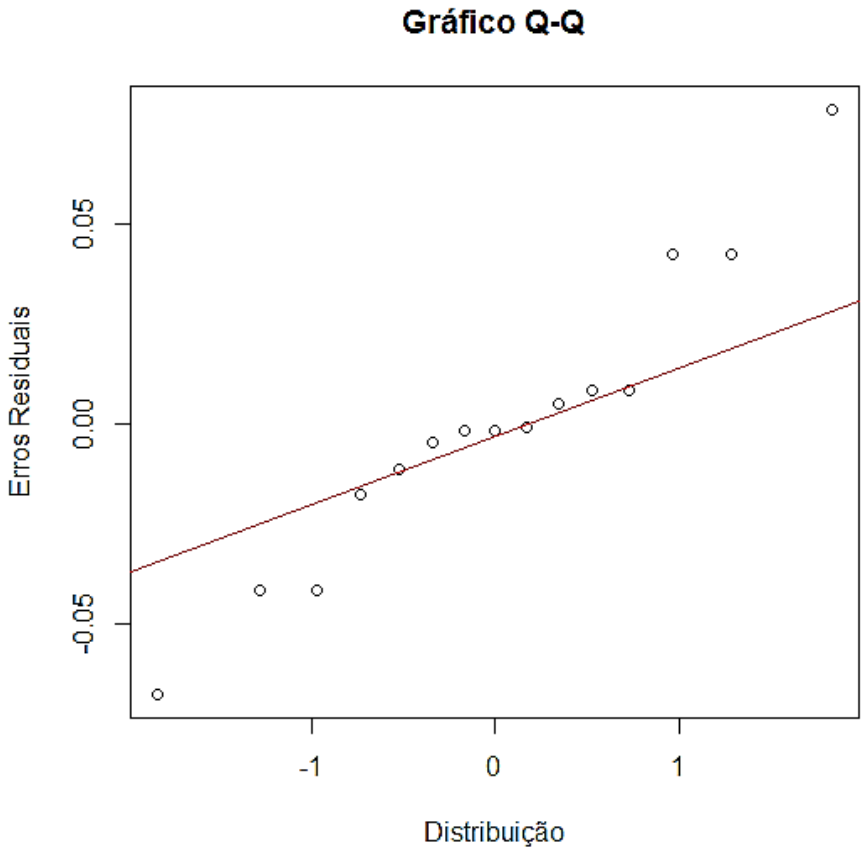
A.2. Testes com Cenário (2) *TrajetosMaiorQueDois*

Esta subseção detalha os resultados oriundos do experimento com o cenário de testes (2) *TrajetosMaiorQueDois*, em que a base de deslocamentos foi filtrada, sendo utilizadas apenas as rotas cujo par <*origem, destino*> foi realizado duas ou mais vezes. As variáveis resposta analisadas pelos testes também foram *precisão, cobertura* e *medida-F*, e, portanto, os resultados serão apresentados e explanados separadamente para cada variável. Todos os testes foram repetidos cinco vezes, assim, as tabelas com os resultados das métricas analisadas apresentarão as médias obtidas após os cinco testes.

Resultados para a Variável Resposta de Precisão

A Tabela 28 apresenta a média dos resultados obtidos para a métrica de *precisão*, após cinco repetições realizadas dos experimentos. As colunas das tabelas apresentam os cinco níveis do *fator F* (modelos de previsão), enquanto as linhas das tabelas representam os três níveis do *bloco B* (percentual de completude da rota a ser testada).

Figura 13 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica medida-F, referente ao cenário de testes (1), da Questão de Pesquisa 1.



Fonte: Elaborada pelo autor.

Tabela 28 – Resultados obtidos para a métrica de Precisão – Preditores x Percentual da rota, para o cenário de teste (2) TrajetosMaiorQueDois.

Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,77	0,50	0,62	0,70	0,19	0,56
50%	0,80	0,58	0,67	0,70	0,19	0,59
85%	0,82	0,61	0,74	0,70	0,19	0,61
Média	**0,80**	**0,56**	**0,68**	**0,70**	**0,19**	**0,59**

Fonte: Elaborada pelo autor.

Uma vez que os resultados das médias obtidos entre os níveis do fator do experimento foram diferentes, o próximo passo foi realizar o teste de hipóteses, com o auxílio da ANOVA. Para isso, os dados apropriados para a análise de variância estão disponíveis na Tabela 29, com a apresentação das médias de *precisão* obtidas, referentes aos níveis do *bloco B* e às alternativas do *fator F*. Além disso, nesta mesma tabela, são apresentados os efeitos dos níveis do bloco e do fator.

Tabela 29 - Adequação dos resultados para aplicação da ANOVA para Precisão, referente ao Cenário (2) TrajetosMaiorQueDois.

Bloco B	***Fator F* – Modelos de Previsão**					**Média bloco**	**Efeito bloco**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM		
15%	0,77	0,50	0,62	0,70	0,19	0,56	-0,03 **(β_1)**
50%	0,80	0,58	0,67	0,70	0,19	0,59	0,00 **(β_2)**
85%	0,82	0,61	0,74	0,70	0,19	0,61	0,03 **(β_3)**
Média alternativas	**0,80 (α_1)**	**0,56 (α_2)**	**0,68 (α_3)**	**0,70 (α_4)**	**0,19 (α_5)**	**0,59**	
Efeito alternativas	**0,21**	**-0,02**	**0,09**	**0,11**	**-0,40**		

Fonte: Elaborada pelo autor.

A Tabela 30 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Até este momento, tem-se os valores da variável resposta de *precisão* (variável *y*), os valores dos efeitos do fator (variável α) e do bloco (variável β) e, a partir deste ponto, estão sendo apresentados os erros (variável e). Com todos os cálculos realizados, a próxima etapa foi calcular a análise de variância.

Tabela 30: Erros experimentais para a Precisão, referentes ao cenário de teste (2) TrajetosMaiorQueDois.

Bloco B	**Erros - *Fator F* – Modelos de Previsão**				
	PPM-Markov	PPM-HMM	PPM	Markov	HMM
15%	0,00	-0,03	-0,03	0,03	0,03
50%	0,00	0,01	-0,01	0,00	0,00
85%	0,00	0,02	0,04	-0,03	-0,03

Fonte: Elaborada pelo autor.

Para o modelo matemático referente à variável resposta *precisão*, tem-se: *SSY = 5,82; SS0 = 5,14, SSB = 0,01; SSA = 0,67; SSE = 0,01; e SST = 0,68*. Com os valores obtidos, obteve-se que o fator influencia, aproximadamente, 98%, o bloco influencia 1% e os erros também 1%, resultado semelhante ao que foi obtido com o cenário (1).

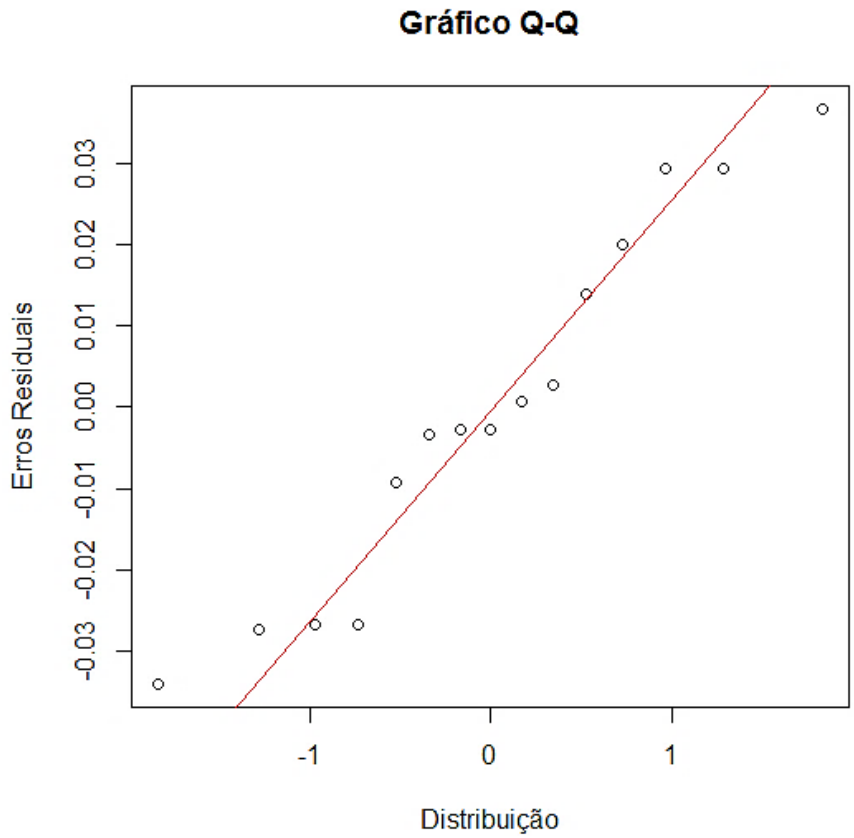
Ao utilizar as Equações (2) e (3) obteve-se, respectivamente, que, para esta análise, o valor de MSA é 0,17 e o de MSE é 0,004. Dividindo-se MSA por MSE, tem-se o cálculo de F, igual a 188. Como a análise realizada nesta pesquisa requer significância de 95%, a tabela F possui um valor de 19,25. Como o cálculo de F pela análise de variância é maior do que o F da tabela (19,25 < 188), é possível refutar, com significância de 95%, a hipótese nula H1-0 da questão de pesquisa ***QP1*** também para o cenário de

testes (2). Isto é, é possível afirmar que, estatisticamente, há diferença nos modelos de previsão avaliados, para prever o destino de uma rota, com respeito à precisão.

Uma vez que a aplicação do teste ANOVA requer que os dados sejam oriundos de uma distribuição normal, deve-se utilizar um teste de normalidade para a verificação dos valores obtidos para a métrica de *precisão*. Ao aplicar o teste de *Shapiro-Wilk* nos erros experimentais, foi possível obter um *valor-p = 0,3695* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de confiança.

Outra abordagem para verificação da normalidade dos dados é o uso de um gráfico Q-Q, disponível na Figura 14. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Figura 14 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica precisão, referente ao cenário de testes (2), da Questão de Pesquisa 1.



Fonte: Elaborada pelo autor.

Resultados para a Variável Resposta de Cobertura

A Tabela 31 apresenta a média dos resultados obtidos para a métrica de *cobertura*, após cinco repetições realizadas dos experimentos. As colunas das tabelas apresentam os cinco níveis do *fator F* (modelos de previsão), enquanto as linhas das tabelas representam os três níveis do *bloco B* (percentual de completude da rota a ser testada).

Tabela 31 – Resultados obtidos para a métrica de Cobertura – Preditores x Percentual da rota, para o cenário de teste (2) TrajetosMaiorQueDois.

Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,86	0,27	0,63	0,80	0,41	0,59
50%	0,89	0,39	0,73	0,80	0,41	0,64
85%	0,91	0,52	0,79	0,80	0,41	0,69
Média	0,89	0,39	0,72	0,80	0,41	**0,64**

Fonte: Elaborada pelo autor.

Uma vez que os resultados das médias obtidos entre os níveis do fator do experimento foram diferentes, o próximo passo foi realizar o teste de hipóteses, com o auxílio da ANOVA. Para isso, os dados apropriados para a análise de variância estão disponíveis na Tabela 32, com a apresentação das médias de *cobertura* obtidas, referentes aos níveis do *bloco B* e às alternativas do *fator F*. Além disso, nesta mesma tabela, são apresentados os efeitos dos níveis do bloco e do fator.

Tabela 32 - Adequação dos resultados para aplicação da ANOVA para Cobertura, referente ao Cenário (2) TrajetosMaiorQueDois.

Bloco B	***Fator F* – Modelos de Previsão**					**Média bloco**	**Efeito bloco**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM		
15%	0,86	0,27	0,63	0,80	0,41	0,59	-0,05 **(β_1)**
50%	0,89	0,39	0,73	0,80	0,41	0,64	0,00 **(β_2)**
85%	0,91	0,52	0,79	0,80	0,41	0,69	0,04 **(β_3)**
Média alternativas	0,89 **(α_1)**	0,39 **(α_2)**	0,72 **(α_3)**	0,80 **(α_4)**	0,41 **(α_5)**	**0,64**	
Efeito alternativas	**0,25**	**-0,25**	**0,08**	**0,16**	**-0,23**		

Fonte: Elaborada pelo autor.

A Tabela 33 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Até este momento, tem-se os valores da variável resposta de *cobertura* (variável *y*), os valores dos efeitos do fator

(variável α) e do bloco (variável β) e, a partir deste ponto, estão sendo apresentados os erros (variável *e*). Com todos os cálculos realizados, a próxima etapa foi calcular a análise de variância.

Tabela 33 - Erros experimentais para a Cobertura, referentes ao cenário de teste (2) TrajetosMaiorQueDois.

Bloco B	**Erros - *Fator F* – Modelos de Previsão**				
	PPM-Markov	PPM-HMM	PPM	Markov	HMM
15%	0,02	-0,08	-0,04	0,05	0,05
50%	0,00	-0,01	0,01	0,00	0,00
85%	-0,02	0,08	0,03	-0,04	-0,04

Fonte: Elaborada pelo autor.

Para o modelo matemático referente à variável resposta *cobertura*, tem-se: *SSY = 6,83; SS0 = 6,17; SSB = 0,02; SSA = 0,62; SSE = 0.02; e SST = 0,66*. Com os valores obtidos, obteve-se que o fator influencia, aproximadamente, 93%, o bloco influencia 3% e os erros influenciam 4%.

Ao utilizar as Equações (2) e (3) obteve-se, respectivamente, que, para esta análise, o valor de MSA é 0,155 e o de MSE é 0,011. Dividindo-se MSA por MSE, tem-se o cálculo de F, igual a 50,7. Como a análise realizada nesta pesquisa requer significância de 95%, a tabela F possui um valor de 19,25. Como o cálculo de F pela análise de variância é maior do que o F da tabela (19,25 < 50,7), é possível refutar, com significância de 95%, a hipótese nula H1-0 da questão de pesquisa ***QP1***, referente à variável resposta de *cobertura*, para o cenário de teste (2). Isto é, pode-se afirmar que, estatisticamente, há diferença nos modelos de previsão avaliados, com respeito à cobertura.

Uma vez que a aplicação do teste ANOVA requer que os dados sejam oriundos de uma distribuição normal, deve-se utilizar um teste de normalidade para a verificação dos valores obtidos para a métrica de *cobertura*. Ao aplicar o teste de *Shapiro-Wilk* nos erros experimentais, foi possível obter um *valor-p = 0,984* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

Outra abordagem para verificação da normalidade dos dados é o uso de um gráfico Q-Q, disponível na Figura 15. Nesta figura, é possível visualizar que a plotagem dos erros

residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Resultados para a Variável Resposta de Medida-F

A Tabela 34 apresenta a média dos resultados obtidos para a métrica de *medida-F*, após cinco repetições realizadas dos experimentos. As colunas das tabelas apresentam os cinco níveis do *fator F* (modelos de previsão), enquanto as linhas das tabelas representam os três níveis do *bloco B* (percentual de completude da rota a ser testada).

Tabela 34 - Resultados obtidos para a métrica de Medida-F – Preditores x Percentual da rota, para o cenário de teste (2) TrajetosMaiorQueDois.

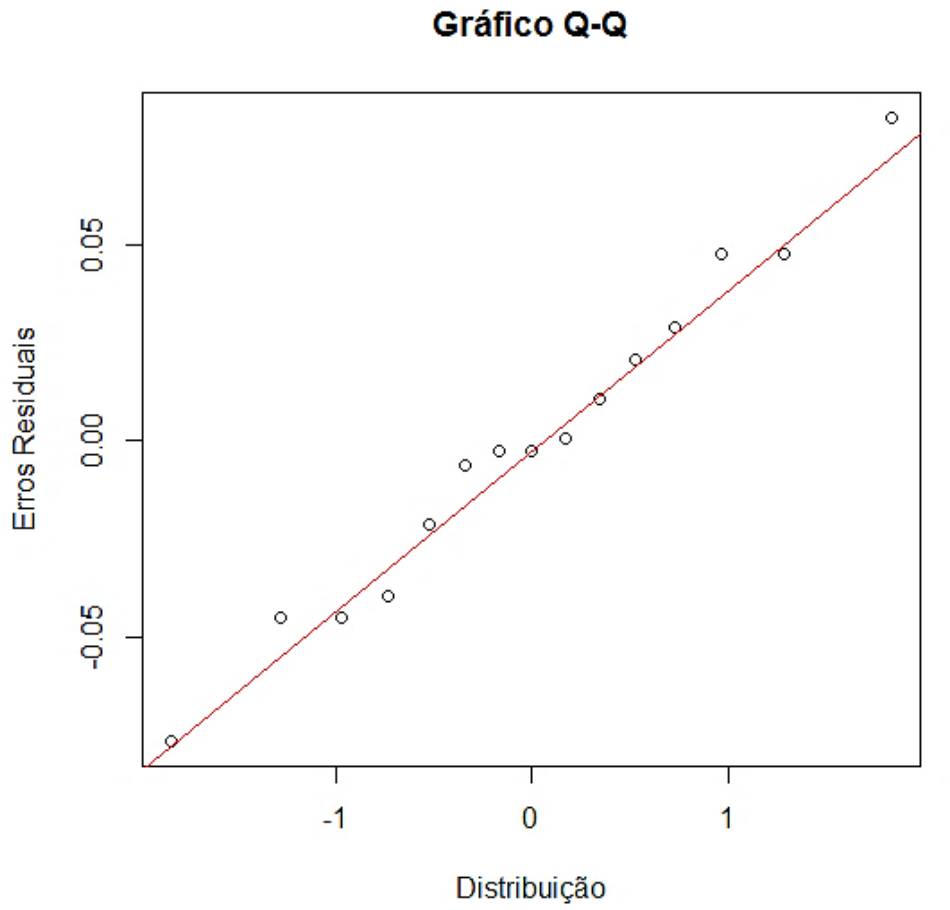
Bloco B	***Fator F* – Modelos de Previsão**					**Média**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM	
15%	0,81	0,35	0,63	0,71	0,24	0,55
50%	0,85	0,47	0,7	0,71	0,24	0,59
85%	0,87	0,56	0,76	0,71	0,24	0,63
Média	0,84	0,46	0,70	0,71	0,24	**0,59**

Fonte: Elaborada pelo autor.

Uma vez que os resultados das médias obtidos entre os níveis do fator do experimento foram diferentes, o próximo passo foi realizar o teste de hipóteses, com o auxílio da ANOVA. Para isso, os dados apropriados para a análise de variância estão disponíveis na Tabela 35, com a apresentação das médias das *medida-F* obtidas, referentes aos níveis do *bloco B* e às alternativas do *fator F*. Além disso, nesta mesma tabela, são apresentados os efeitos dos níveis do bloco e do fator.

A Tabela 36 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Até este momento, tem-se os valores da variável resposta de *medida-F* (variável *y*), os valores dos efeitos do fator (variável α) e do bloco (variável β) e, a partir deste ponto, estão sendo apresentados os erros (variável *e*). Com todos os cálculos realizados, a próxima etapa foi calcular a análise de variância.

Figura 15 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica cobertura, referente ao cenário de testes (2), da Questão de Pesquisa 1.



Fonte: Elaborada pelo autor.

Tabela 35 - Adequação dos resultados para aplicação da ANOVA para Medida-F, referente ao Cenário (2) TrajetosMaiorQueDois.

Bloco B	***Fator F* – Modelos de Previsão**					**Média bloco**	**Efeito bloco**
	PPM-Markov	PPM-HMM	PPM	Markov	HMM		
15%	0,81	0,35	0,63	0,71	0,24	0,55	-0,04 (β_1)
50%	0,85	0,47	0,7	0,71	0,24	0,59	0,00 (β_2)
85%	0,87	0,56	0,76	0,71	0,24	0,63	0,04 (β_3)
Média alternativas	0,84 (α_1)	0,46 (α_2)	0,70 (α_3)	0,71 (α_4)	0,24 (α_5)	**0,60**	
Efeito alternativas	**0,25**	**-0,13**	**0,11**	**0,12**	**-0,25**		

Fonte: Elaborada pelo autor.

Tabela 36 - Erros experimentais para a Medida-F, referentes ao cenário de teste (2) TrajetosMaiorQueDois.

Bloco B	**Erros - *Fator F* – Modelos de Previsão**				
	PPM-Markov	PPM-HMM	PPM	Markov	HMM
15%	0,01	-0,07	-0,02	0,04	0,04
50%	0,00	0,01	0,00	0,00	0,00
85%	-0,01	0,06	0,03	-0,04	-0,04

Fonte: Elaborada pelo autor.

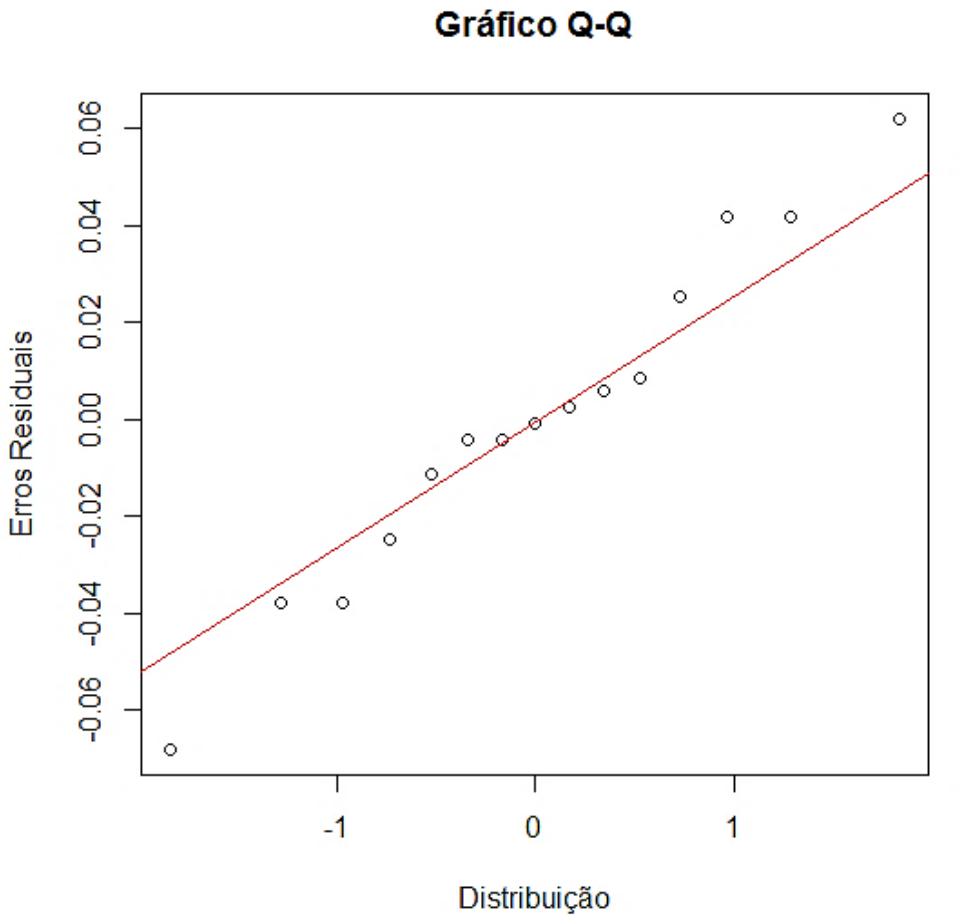
Para o modelo matemático referente à variável resposta *medida-F*, tem-se: *SSY = 5,94; SS0 = 5,22; SSB = 0.02; SSA = 0,69; SSE = 002; e SST = 0,72*. Com os valores obtidos, obteve-se que o fator influencia, aproximadamente, 95,5%, o bloco influencia 2,2% e os erros 2,3%.

Ao utilizar as Equações (2) e (3) obteve-se, respectivamente, que, para esta análise, o valor de MSA é 0,172 e o de MSE é 0,002. Dividindo-se MSA por MSE, tem-se o cálculo de F, igual a 83,8. Como a análise realizada neste trabalho requer significância de 95%, a tabela F possui um valor de 19,25. Como o cálculo de F pela análise de variância é maior do que o F da tabela (19,25 < 83,4), é possível refutar, com significância de 95%, a hipótese nula H1-0 da questão de pesquisa ***QP1***, referente à variável resposta de *medida-F*. Isto é, pode-se afirmar que, estatisticamente, há diferença nos modelos de previsão avaliados, para prever o destino de uma rota, também para o cenário de teste (2), com trajetos cujo par <*origem*, *destino*> foram realizados duas ou mais vezes, com respeito à medida-F.

Uma vez que a aplicação do teste ANOVA requer que os dados sejam oriundos de uma distribuição normal, deve-se utilizar um teste de normalidade para a verificação dos valores obtidos para a métrica de *medida-F*. Ao aplicar o teste de *Shapiro-Wilk* nos erros experimentais, foi possível obter um *valor-p = 0,9415* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

Outra abordagem para verificação da normalidade dos dados é o uso de um gráfico Q-Q, disponível na Figura 16. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Figura 16 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica medida-F, referente ao cenário de testes (2), da Questão de Pesquisa 1.



Fonte: Elaborada pelo autor.

APÊNDICE B – Detalhamento dos Resultados Obtidos com os Experimentos para a Questão de Pesquisa 2

Neste apêndice, serão detalhados os resultados obtidos com os experimentos realizados para a questão de pesquisa **QP2**, apresentada na Seção 1.2, que trata do seguinte problema.

QP2 – *Existe diferença no resultado da previsão de trajetórias em uma base de dados com rotas que foram realizadas mais frequentemente (isto é, onde o par <origem, destino> foi realizado pelo menos duas vezes) versus uma base de dados que possui mais rotas que foram realizadas uma única vez (isto é, onde o par <origem, destino> foi realizado apenas uma vez) para os modelos de previsão baseados em PPM?*

H2-0: Não há diferença no resultado da previsão de trajetórias em usar bases com rotas frequentes *versus* bases com rotas que contenham muitos trajetos realizados apenas uma vez para previsão de trajetória, referentes às métricas estatísticas precisão, cobertura e medida-F.

H2-1: Há diferença no resultado da previsão de trajetórias em usar bases com rotas frequentes *versus* bases com rotas que contenham muitos trajetos realizados apenas uma vez para previsão de trajetória, referentes às métricas estatísticas precisão, cobertura e medida-F.

Conforme já exposto na Seção 5.3.1, foi adotado o *desgin fatorial* (com um único fator) *com blocagem*, em que o *fator F* refere-se ao *uso da base de dados* (com dois níveis) e o *bloco B* representa o *percentual de completude das rotas a serem testadas* (com três níveis). Os dois níveis do *fator F* são *TodosTrajetos* e *TrajetosMaiorQueDois*, e os três níveis do *bloco B* são 15%, 50% e 85%. Os testes foram realizados separadamente para cada um dos três modelos de previsão inovadores propostos nesta tese, resultando, portanto, em três subseções neste Apêndice: uma para contemplar o modelo de previsão PPM-Markov; outra para contemplar o modelo de previsão PPM-HMM; e outra para contemplar o modelo de previsão PPM puro, em que não há combinação do PPM com outro método. As próximas três subseções apresentam os resultados estatísticos, inclusive com aplicação do teste de *Análise de Variância* (ANOVA) e do teste de normalidade.

Como na análise da **QP1** (disponível no APÊNDICE A deste documento) os resultados foram detalhados de maneira a tentar facilitar o entendimento, os resultados apresentados para a **QP2** serão sumarizados para fins de melhoria da legibilidade. Isto é, os valores dos efeitos gerados pelo fator e pelo bloco serão sumarizados e apresentados em tabelas, ao invés de serem relatados textualmente. Além disso, os resultados obtidos para as variáveis das *Somas dos Quadrados* (SSY, SSA, SSB, SSE, SS0 e SST) e a *Média dos Quadrados para o Fator* (MSA) e *para o Erro* (MSE) estarão dispostos em tabela. Estas tabelas estarão organizadas separadamente conforme as variáveis resposta analisadas, que são, assim como na **QP1**, *precisão, cobertura* e *medida-F*.

B.1. Testes com o Modelo de Previsão PPM-Markov

Esta seção apresenta os experimentos agrupados pelo modelo de previsão PPM com Markov. Para o experimento, o *fator F* possui os níveis referentes ao uso da base de dados, cujos valores podem ser a base *TodosTrajetos* e a base *TrajetosMaiorQueDois*, enquanto o *bloco B* refere-se ao percentual de completude da rota a ser testada (15%, 50% e 85%). As variáveis resposta analisadas pelos testes foram *precisão, cobertura* e *medida-F*.

Como são três variáveis resposta a serem analisadas separadamente, foi criada uma subseção para cada uma delas, e elaborada uma quarta subseção destinada apenas aos comentários dos resultados obtidos.

Resultados para a Variável Resposta de Precisão

A Tabela 37 apresenta os resultados obtidos após a realização do experimento. A primeira coluna da tabela é destinada aos níveis do *bloco B* (exceto as duas últimas linhas), enquanto a segunda coluna apresenta os resultados das médias obtidos para a *precisão* referente aos dois níveis do *fator F*, após cinco repetições. A terceira e a quarta colunas apresentam, respectivamente, as médias obtidas por bloco e o efeito gerado pelo bloco. A penúltima linha da tabela apresenta as médias de cada uma das duas alternativas do fator, enquanto a última linha apresenta o efeito gerado pelas alternativas do *fator F*.

Tabela 37 - Resultados obtidos para a métrica de Precisão – Base de dados x Percentual da rota, para o modelo PPM-Markov.

Bloco B	***Fator F* – Base de dados**		**Média bloco**	**Efeito bloco**
	TodosTrajetos	*TrajetosMaiorQueDois*		
15%	0,63	0,77	0,70	-0,03 **(β_1)**
50%	0,67	0,80	0,74	0,01 **(β_2)**
85%	0,66	0,82	0,74	0,01 **(β_3)**
Média alternativas	**0,65 (α_1)**	**0,80 (α_2)**	**0,73**	
Efeito alternativas	**-0,07**	**0,07**		

Fonte: Elaborada pelo autor.

A Tabela 38 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Já a Tabela 39 apresenta os valores das *Somas dos Quadrados* (*SSY, SS0, SSB, SSA, SSE* e *SST*), além dos valores de MSA e de MSE. Ainda nesta tabela, os percentuais de influência do fator, do bloco e dos erros, bem como o cálculo de F para o experimento realizado e o F crítico (F_c) da tabela F também são apresentados.

Tabela 38 - Erros experimentais para a Precisão, referentes ao preditor PPM-Markov.

Bloco B	***Fator F* – Base de dados**	
	TodosTrajetos	*TrajetosMaiorQueDois*
15%	0,00	0,00
50%	0,01	-0,01
85%	-0,01	0,01

Fonte: Elaborada pelo autor.

Tabela 39 - Resultados estatísticos consolidados para a Precisão, referentes ao preditor PPM-Markov.

Resultados Estatísticos do Modelo PPM-Markov – *Precisão*							
SSY:	3,19	*SS0:*	3,15	*SSE:*	0,0002	*MSA:*	0,031
SSA:	0,031	*SSB:*	0,002	*SST:*	0,03	*MSE:*	0,0001
Percentual de Influência do		*Fator:*	98,6%	*Valor De F:*	264	*F_c da tabela F (95%)*	18,51
		Bloco:	0,8%				
		Erro:	0,6%				

Fonte: Elaborada pelo autor.

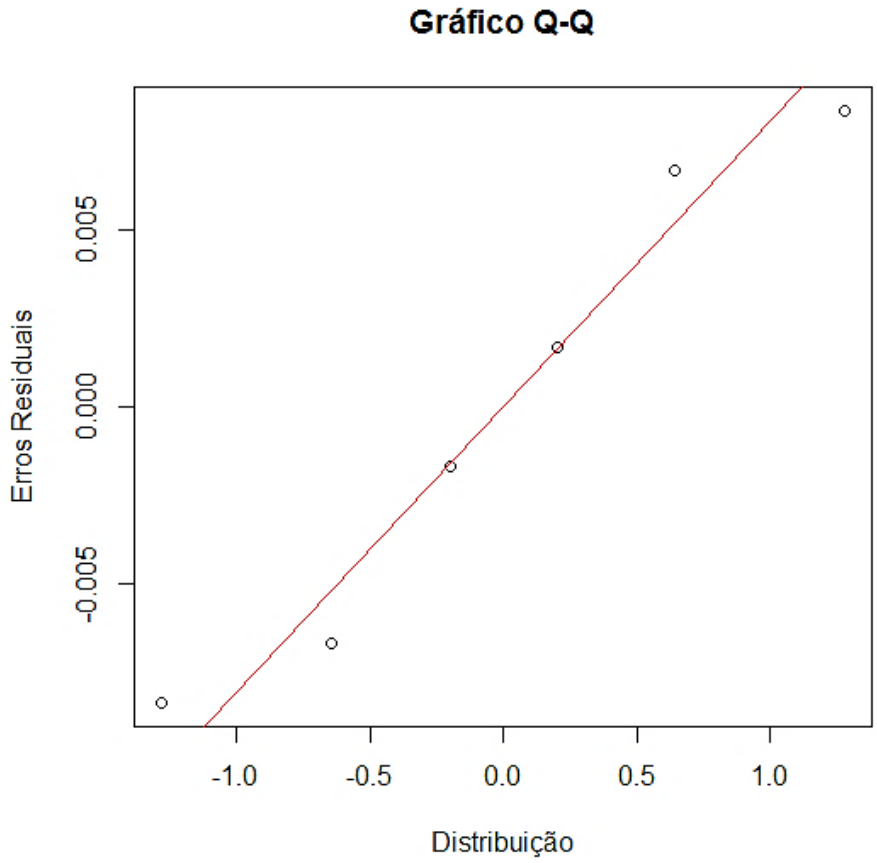
Como a análise realizada neste trabalho requer significância de 95%, a tabela F possui um valor de 18,51. Uma vez que o cálculo de F obtido pela análise de variância é maior do que o F da tabela (18,51 < 264), é possível refutar, com significância de 95%, a

hipótese nula H2-0 da questão de pesquisa ***QP2***, para a métrica de *precisão* referente ao modelo de previsão PPM-Markov. Isto é, é possível afirmar que, estatisticamente, há diferença em utilizar a base de dados do cenário *TodosTrajetos* comparada com a base do cenário *TrajetosMaiorQueDois*, com respeito à *precisão*.

Para a verificação de que os valores utilizados no teste ANOVA, referentes à métrica de *precisão*, são oriundos de uma distribuição normal, foi aplicado o teste de normalidade *Shapiro-Wilk* nos erros experimentais. Da aplicação deste teste, foi possível obter um *valor-p = 0,602* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

O gráfico Q-Q, apresentado na Figura 17, representa outra abordagem para verificação da normalidade. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Figura 17 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica *Precisão*, referente ao uso do modelo de previsão PPM-Markov, para a Questão de Pesquisa 2.



Fonte: Elaborada pelo autor.

Resultados para a Variável Resposta de Cobertura

A Tabela 40 apresenta os resultados obtidos após a realização do experimento. A primeira coluna da tabela é destinada aos níveis do *bloco B* (exceto as duas últimas linhas), enquanto a segunda coluna apresenta os resultados das médias obtidos para a *cobertura* referente aos dois níveis do *fator F*, após cinco repetições. A terceira e a quarta colunas apresentam, respectivamente, as médias obtidas por bloco e o efeito gerado pelo bloco. A penúltima linha da tabela apresenta as médias de cada uma das duas alternativas do fator, enquanto a última linha apresenta o efeito gerado pelas alternativas do *fator F*.

Tabela 40 - Resultados obtidos para a métrica de Cobertura – Base de dados x Percentual da rota, para o modelo PPM-Markov.

Bloco B	***Fator F* – Base de dados**		**Média bloco**	**Efeito bloco**
	TodosTrajetos	*TrajetosMaiorQueDois*		
15%	0,32	0,86	0,59	-0,04 **(β_1)**
50%	0,38	0,89	0,64	0,00 **(β_2)**
85%	0,44	0,91	0,68	0,04 **(β_3)**
Média alternativas	**0,39 (α_1)**	**0,89 (α_2)**	**0,63**	
Efeito alternativas	**-0,25**	**0,25**		

Fonte: Elaborada pelo autor.

A Tabela 41 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Já a Tabela 42 apresenta os valores das *Somas dos Quadrados* (*SSY, SS0, SSB, SSA, SSE* e *SST*), além dos valores de MSA e de MSE. Ainda nesta tabela, os percentuais de influência do fator, do bloco e dos erros, bem como o cálculo de F para o experimento realizado e o F crítico (F_c) da tabela F também são apresentados.

Tabela 41 - Erros experimentais para a Cobertura, referentes ao preditor PPM-Markov.

Bloco B	***Fator F* – Base de dados**	
	TodosTrajetos	*TrajetosMaiorQueDois*
15%	-0,02	0,02
50%	0,00	0,00
85%	0,02	-0,02

Fonte: Elaborada pelo autor.

Tabela 42 - Resultados estatísticos consolidados para a Cobertura, referentes ao preditor PPM-Markov.

Resultados Estatísticos do Modelo PPM-Markov – *Cobertura*							
SSY:	2,80	*SS0:*	2,41	*SSE:*	0,001	*MSA:*	0,39
SSA:	0,39	*SSB:*	0,007	*SST:*	0,39	*MSE:*	0,001
Percentual de Influência do		*Fator:*	98%	*Valor De F:*	624	*F_c da tabela F (95%)*	18,51
		Bloco:	1,8%				
		Erro:	0,2%				

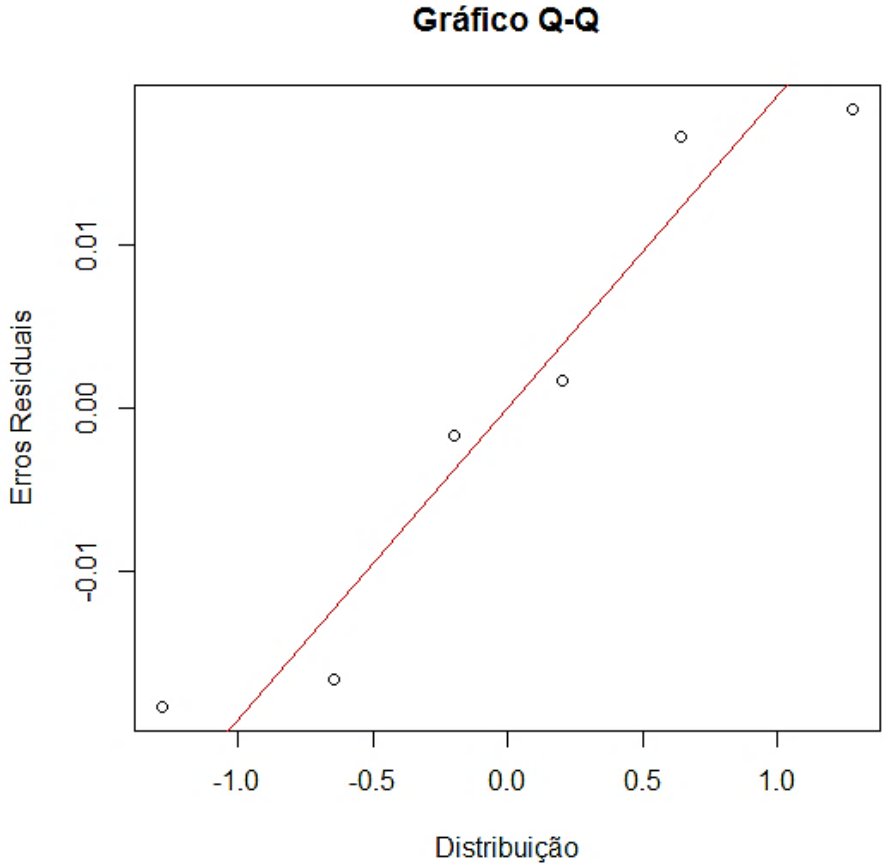
Fonte: Elaborada pelo autor.

Como a análise realizada neste trabalho requer significância de 95%, a tabela F possui um valor de 18,51. Uma vez que o cálculo de F obtido pela análise de variância é maior do que o F da tabela (18,51 < 624), é possível refutar, com significância de 95%, a hipótese nula H2-0 da questão de pesquisa ***QP2***, para a métrica de *cobertura* referente ao modelo de previsão PPM-Markov. Isto é, é possível afirmar que, estatisticamente, há diferença em utilizar a base de dados do cenário *TodosTrajetos* comparada com a base do cenário *TrajetosMaiorQueDois*, com respeito à *cobertura*.

Para a verificação de que os valores utilizados no teste ANOVA, referentes à métrica de *cobertura*, são oriundos de uma distribuição normal, foi aplicado o teste de normalidade *Shapiro-Wilk* nos erros experimentais. Da aplicação deste teste, foi possível obter um *valor-p = 0,3461* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

O gráfico Q-Q, apresentado na Figura 18, representa outra abordagem para verificação da normalidade. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Figura 18 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica *Cobertura*, referente ao uso do modelo de previsão PPM-Markov, para a Questão de Pesquisa 2.



Fonte: Elaborada pelo autor.

Resultados para a Variável Resposta de Medida-F

A Tabela 43 apresenta os resultados obtidos após a realização do experimento. A primeira coluna da tabela é destinada aos níveis do *bloco B* (exceto as duas últimas linhas), enquanto a segunda coluna apresenta os resultados das médias obtidos para a *medida-F* referente aos dois níveis do *fator F*, após cinco repetições. A terceira e a quarta colunas apresentam, respectivamente, as médias obtidas por bloco e o efeito gerado pelo bloco. A penúltima linha da tabela apresenta as médias de cada uma das duas alternativas do fator, enquanto a última linha apresenta o efeito gerado pelas alternativas do *fator F*.

A Tabela 44 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Já a Tabela 45 apresenta os valores das *Somas dos Quadrados* (*SSY, SS0, SSB, SSA, SSE* e *SST*), além dos valores de MSA e de MSE. Ainda nesta tabela, os percentuais de influência do fator, do bloco e dos erros, bem como o cálculo de F para o experimento realizado e o F crítico (F_c) da tabela F também são apresentados.

Tabela 43 - Resultados obtidos para a métrica de Medida-F – Base de dados x Percentual da rota, para o modelo PPM-Markov.

Bloco B	***Fator F* – Base de dados**		**Média bloco**	**Efeito bloco**
	TodosTrajetos	*TrajetosMaiorQueDois*		
15%	0,42	0,81	0,62	-0,05 **(β_1)**
50%	0,49	0,85	0,67	0,01 **(β_2)**
85%	0,53	0,87	0,70	0,04 **(β_3)**
Média alternativas	**0,48 (α_1)**	**0,84 (α_2)**	**0,66**	
Efeito alternativas	**-0,18**	**0,18**		

Fonte: Elaborada pelo autor.

Tabela 44 - Erros experimentais para a Medida-F, referentes ao preditor PPM-Markov.

Bloco B	***Fator F* – Base de dados**	
	TodosTrajetos	*TrajetosMaiorQueDois*
15%	-0,01	0,01
50%	0,00	0,00
85%	0,01	-0,01

Fonte: Elaborada pelo autor.

Tabela 45 - Resultados estatísticos consolidados para a Medida-F, referentes ao preditor PPM-Markov.

Resultados Estatísticos do Modelo PPM-Markov – *Medida-F*							
SSY:	2,83	*SS0:*	2,63	*SSE:*	0,0006	*MSA:*	0,2
SSA:	0,2	*SSB:*	0,007	*SST:*	0,21	*MSE:*	0,0006
Percentual de Influência do		*Fator:*	96,3%	*Valor De F:*	625	*F_c da tabela F (95%)*	18,51
		Bloco:	2,4%				
		Erro:	1,3%				

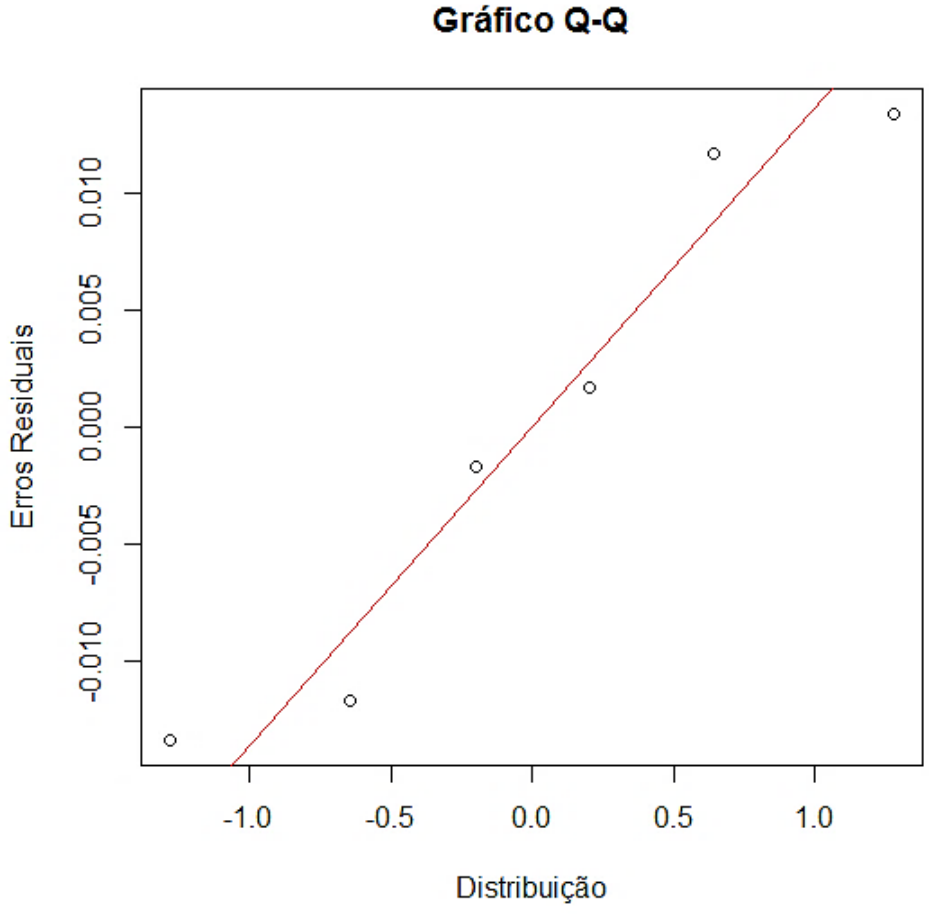
Fonte: Elaborada pelo autor.

Como a análise realizada neste trabalho requer significância de 95%, a tabela F possui um valor de 18,51. Uma vez que o cálculo de F obtido pela análise de variância é maior do que o F da tabela (18,51 < 625), é possível refutar, com significância de 95%, a hipótese nula H2-0 da questão de pesquisa ***QP2***, para a métrica de *medida-F* referente ao modelo de previsão PPM-Markov. Isto é, é possível afirmar que, estatisticamente, há diferença em utilizar a base de dados do cenário *TodosTrajetos* comparada com a base do cenário *TrajetosMaiorQueDois*, com respeito à *medida-F*.

Para a verificação de que os valores utilizados no teste ANOVA, referentes à métrica de *medida-F*, são oriundos de uma distribuição normal, foi aplicado o teste de normalidade *Shapiro-Wilk* nos erros experimentais. Da aplicação deste teste, foi possível obter um *valor-p = 0,428* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

O gráfico Q-Q, apresentado na Figura 19, representa outra abordagem para verificação da normalidade. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Figura 19 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica *Medida-F*, referente ao uso do modelo de previsão PPM-Markov, para a Questão de Pesquisa 2.



Fonte: Elaborada pelo autor.

B.2. Testes com o Modelo de Previsão PPM-HMM

Esta seção apresenta os experimentos agrupados pelo modelo de previsão PPM com Cadeias Ocultas de Markov (HMM). Para o experimento, o *fator F* possui os níveis referentes ao uso da base de dados, cujos valores podem ser a base *TodosTrajetos* e a base *TrajetosMaiorQueDois*, enquanto o *bloco B* refere-se ao percentual de completude da rota a ser testada (15%, 50% e 85%). As variáveis resposta analisadas pelos testes foram *precisão, cobertura* e *medida-F*.

Como são três variáveis resposta a serem analisadas separadamente, foi criada uma subseção para cada uma delas, e elaborada uma quarta subseção destinada apenas aos comentários dos resultados obtidos.

Resultados para a Variável Resposta de Precisão

A Tabela 46 apresenta os resultados obtidos após a realização do experimento. A primeira coluna da tabela é destinada aos níveis do *bloco B* (exceto as duas últimas linhas), enquanto a segunda coluna apresenta os resultados das médias obtidos para a *precisão* referente aos dois níveis do *fator F*, após cinco repetições. A terceira e a quarta colunas apresentam, respectivamente, as médias obtidas por bloco e o efeito gerado pelo bloco. A penúltima linha da tabela apresenta as médias de cada uma das duas alternativas do fator, enquanto a última linha apresenta o efeito gerado pelas alternativas do *fator F*.

Tabela 46 - Resultados obtidos para a métrica de Precisão – Base de dados x Percentual da rota, para o modelo PPM-HMM.

Bloco B	***Fator F* – Base de dados**		**Média bloco**	**Efeito bloco**
	TodosTrajetos	*TrajetosMaiorQueDois*		
15%	0,44	0,5	0,47	-0,06 **(β_1)**
50%	0,50	0,58	0,58	0,01 **(β_2)**
85%	0,54	0,61	0,61	0,05 **(β_3)**
Média alternativas	**0,49 (α_1)**	**0,56 (α_2)**	**0,67**	
Efeito alternativas	**-0,04**	**0,04**		

Fonte: Elaborada pelo autor.

A Tabela 47 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Já a Tabela 48 apresenta os valores das *Somas dos Quadrados* (*SSY, SS0, SSB, SSA, SSE* e *SST*), além dos valores de MSA e de MSE. Ainda nesta tabela, os percentuais de influência do fator, do bloco e

dos erros, bem como o cálculo de F para o experimento realizado e o F crítico (F_c) da tabela F também são apresentados.

Tabela 47 - Erros experimentais para a Precisão, referentes ao preditor PPM-HMM.

Bloco B	***Fator F* – Base de dados**	
	TodosTrajetos	*TrajetosMaiorQueDois*
15%	0,01	-0,01
50%	-0,01	0,00
85%	0,00	0,00

Fonte: Elaborada pelo autor.

Tabela 48 - Resultados estatísticos consolidados para a Precisão, referentes ao preditor PPM-HMM.

Resultados Estatísticos do Modelo PPM-HMM – *Precisão*							
SSY:	1,69	*SS0:*	1,67	*SSE:*	0,00	*MSA:*	0,0074
SSA:	0,0074	*SSB:*	0,0114	*SST:*	0,019	*MSE:*	0,00
Percentual de Influência do		*Fator:*	39%	*Valor De F:*	147	*F_c da tabela F (95%)*	18,51
		Bloco:	61%				
		Erro:	0%				

Fonte: Elaborada pelo autor.

Como a análise realizada neste trabalho requer significância de 95%, a tabela F possui um valor de 18,51. Uma vez que o cálculo de F obtido pela análise de variância é maior do que o F da tabela (18,51 < 147), é possível refutar, com significância de 95%, a hipótese nula H2-0 da questão de pesquisa ***QP2***, para a métrica de *precisão* referente ao modelo de previsão PPM-HMM. Isto é, é possível afirmar que, estatisticamente, há diferença em utilizar a base de dados do cenário *TodosTrajetos* comparada com a base do cenário *TrajetosMaiorQueDois*, com respeito à *precisão*.

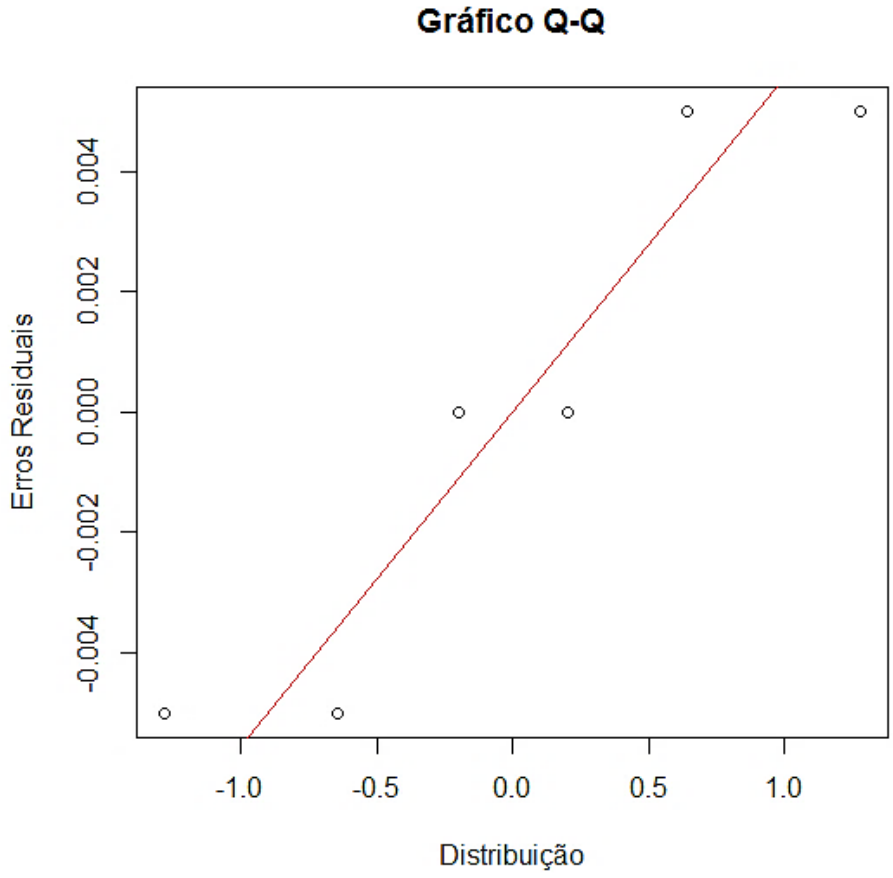
Com relação aos testes, é importante destacar que o bloco obteve 61% de influência nos resultados, valor superior aos 39% de influência que foram obtidos pelo fator. Portanto, é possível refutar a hipótese nula da **QP2**, com relação à métrica de precisão, porém, a maior influência nos resultados foi do *bloco B*, referente ao percentual de completude das rotas a serem testadas.

Para a verificação de que os valores utilizados no teste ANOVA, referentes à métrica de *precisão*, são oriundos de uma distribuição normal, foi aplicado o teste de normalidade *Shapiro-Wilk* nos erros experimentais. Da aplicação deste teste, foi possível obter um *valor-p = 0,167* (> 0,05), resultando, portanto, na afirmação de que não é

possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

O gráfico Q-Q, apresentado na Figura 20, representa outra abordagem para verificação da normalidade. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Figura 20 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica *Precisão*, referente ao uso do modelo de previsão PPM-HMM, para a Questão de Pesquisa 2.



Fonte: Elaborada pelo autor.

Resultados para a Variável Resposta de Cobertura

A Tabela 49 apresenta os resultados obtidos após a realização do experimento. A primeira coluna da tabela é destinada aos níveis do *bloco B* (exceto as duas últimas linhas), enquanto a segunda coluna apresenta os resultados das médias obtidos para a *cobertura* referente aos dois níveis do *fator F*, após cinco repetições. A terceira e a quarta

colunas apresentam, respectivamente, as médias obtidas por bloco e o efeito gerado pelo bloco. A penúltima linha da tabela apresenta as médias de cada uma das duas alternativas do fator, enquanto a última linha apresenta o efeito gerado pelas alternativas do *fator F*.

Tabela 49 - Resultados obtidos para a métrica de Cobertura – Base de dados x Percentual da rota, para o modelo PPM-HMM.

Bloco B	***Fator F* – Base de dados**		**Média bloco**	**Efeito bloco**
	TodosTrajetos	*TrajetosMaiorQueDois*		
15%	0,24	0,27	0,26	-0,09 **(β_1)**
50%	0,28	0,39	0,34	-0,01 **(β_2)**
85%	0,39	0,52	0,46	0,11 **(β_3)**
Média alternativas	**0,30 (α_1)**	**0,39 (α_2)**	**0,35**	
Efeito alternativas	**-0,05**	**0,05**		

Fonte: Elaborada pelo autor.

A Tabela 50 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Já a Tabela 51 apresenta os valores das *Somas dos Quadrados* (*SSY, SS0, SSB, SSA, SSE* e *SST*), além dos valores de MSA e de MSE. Ainda nesta tabela, os percentuais de influência do fator, do bloco e dos erros, bem como o cálculo de F para o experimento realizado e o F crítico (F_c) da tabela F também são apresentados.

Como a análise realizada neste trabalho requer significância de 95%, a tabela F possui um valor de 18,51. Uma vez que o cálculo de F obtido pela análise de variância é menor do que o F da tabela (18,51 > 8,68), não é possível refutar, com significância de 95%, a hipótese nula H2-0 da questão de pesquisa ***QP2***, para a métrica de *cobertura* referente ao modelo de previsão PPM-HMM. Isto é, não é possível afirmar que, estatisticamente, há diferença em utilizar a base de dados do cenário *TodosTrajetos* comparada com a base do cenário *TrajetosMaiorQueDois*, com respeito à *cobertura*.

Tabela 50 - Erros experimentais para a Cobertura, referentes ao preditor PPM-HMM.

Bloco B	***Fator F* – Base de dados**	
	TodosTrajetos	*TrajetosMaiorQueDois*
15%	0,03	-0,03
50%	-0,01	0,01
85%	-0,02	0,02

Fonte: Elaborada pelo autor.

Tabela 51 - Resultados estatísticos consolidados para a Cobertura, referentes ao preditor PPM-HMM.

Resultados Estatísticos do Modelo PPM-HMM – *Cobertura*							
SSY:	0,78	*SS0:*	10,73	*SSE:*	0,0028	*MSA:*	0,123
SSA:	0,012	*SSB:*	0,041	*SST:*	0,5548	*MSE:*	0,0014
Percentual de Influência do		*Fator:*	22%	*Valor De F:*	8,68	*F_c da tabela F (95%)*	18,51
		Bloco:	73%				
		Erro:	5%				

Fonte: Elaborada pelo autor.

Diferentemente do que foi obtido para a métrica de *precisão*, para a variável resposta de *cobertura*, o fator obteve maior influência na obtenção dos resultados estatísticos, com 89,5%. Já o bloco, teve influência de 9,6% no nos valores obtidos pelo experimento.

Para a verificação de que os valores utilizados no teste ANOVA, referentes à métrica de *cobertura*, são oriundos de uma distribuição normal, foi aplicado o teste de normalidade *Shapiro-Wilk* nos erros experimentais. Da aplicação deste teste, foi possível obter um *valor-p = 0,7393* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

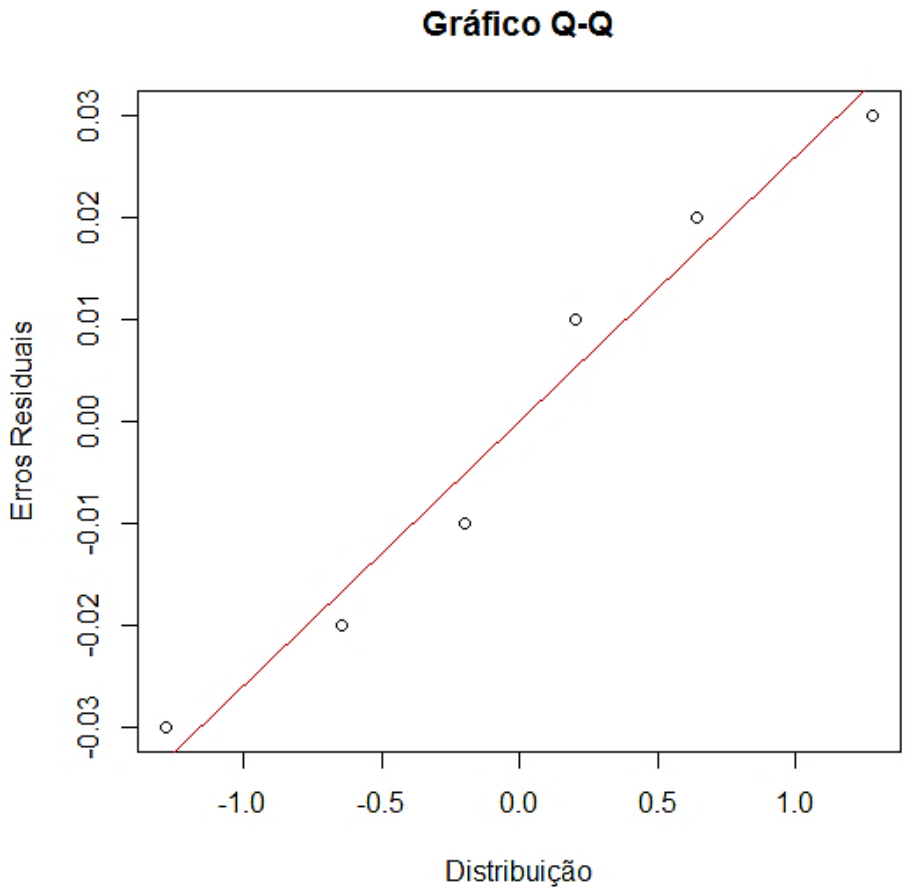
O gráfico Q-Q, apresentado na Figura 21, representa outra abordagem para verificação da normalidade. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Resultados para a Variável Resposta de Medida-F

A Tabela 52 apresenta os resultados obtidos após a realização do experimento. A primeira coluna da tabela é destinada aos níveis do *bloco B* (exceto as duas últimas linhas), enquanto a segunda coluna apresenta os resultados das médias obtidos para a *medida-F* referente aos dois níveis do *fator F*, após cinco repetições. A terceira e a quarta colunas apresentam, respectivamente, as médias obtidas por bloco e o efeito gerado pelo

bloco. A penúltima linha da tabela apresenta as médias de cada uma das duas alternativas do fator, enquanto a última linha apresenta o efeito gerado pelas alternativas do *fator F*.

Figura 21 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica *Cobertura*, referente ao uso do modelo de previsão PPM-HMM, para a Questão de Pesquisa 2.



Fonte: Elaborada pelo autor.

Tabela 52 - Resultados obtidos para a métrica de Medida-F – Base de dados x Percentual da rota, para o modelo PPM-HMM.

Bloco B	***Fator F* – Base de dados**		**Média bloco**	**Efeito bloco**
	TodosTrajetos	*TrajetosMaiorQueDois*		
15%	0,31	0,35	0,33	-0,08 **(β_1)**
50%	0,36	0,47	0,42	0,01 **(β_2)**
85%	0,39	0,56	0,48	0,07 **(β_3)**
Média alternativas	**0,35 (α_1)**	**0,46 (α_2)**	**0,41**	
Efeito alternativas	**-0,05**	**0,05**		

Fonte: Elaborada pelo autor.

A Tabela 53 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Já a Tabela 54 apresenta os valores das *Somas dos Quadrados* (*SSY, SS0, SSB, SSA, SSE* e *SST*), além dos valores

de MSA e de MSE. Ainda nesta tabela, os percentuais de influência do fator, do bloco e dos erros, bem como o cálculo de F para o experimento realizado e o F crítico (F_c) da tabela F também são apresentados.

Tabela 53 - Erros experimentais para a Medida-F, referentes ao preditor PPM-HMM.

Bloco B	***Fator F* – Base de dados**	
	TodosTrajetos	*TrajetosMaiorQueDois*
15%	0,03	-0,03
50%	0,00	0,00
85%	-0,03	0,03

Fonte: Elaborada pelo autor.

Tabela 54 - Resultados estatísticos consolidados para a Medida-F, referentes ao preditor PPM-HMM.

Resultados Estatísticos do Modelo PPM-Markov – *Medida-F*							
SSY:	1,03	*SS0:*	0,99	*SSE:*	0,004	*MSA:*	0,017
SSA:	0,017	*SSB:*	0,021	*SST:*	0,043	*MSE:*	0,0106
Percentual de Influência do		*Fator:*	40%	*Valor De F:*	8	*F_c da tabela F (95%)*	18,51
		Bloco:	50%				
		Erro:	10%				

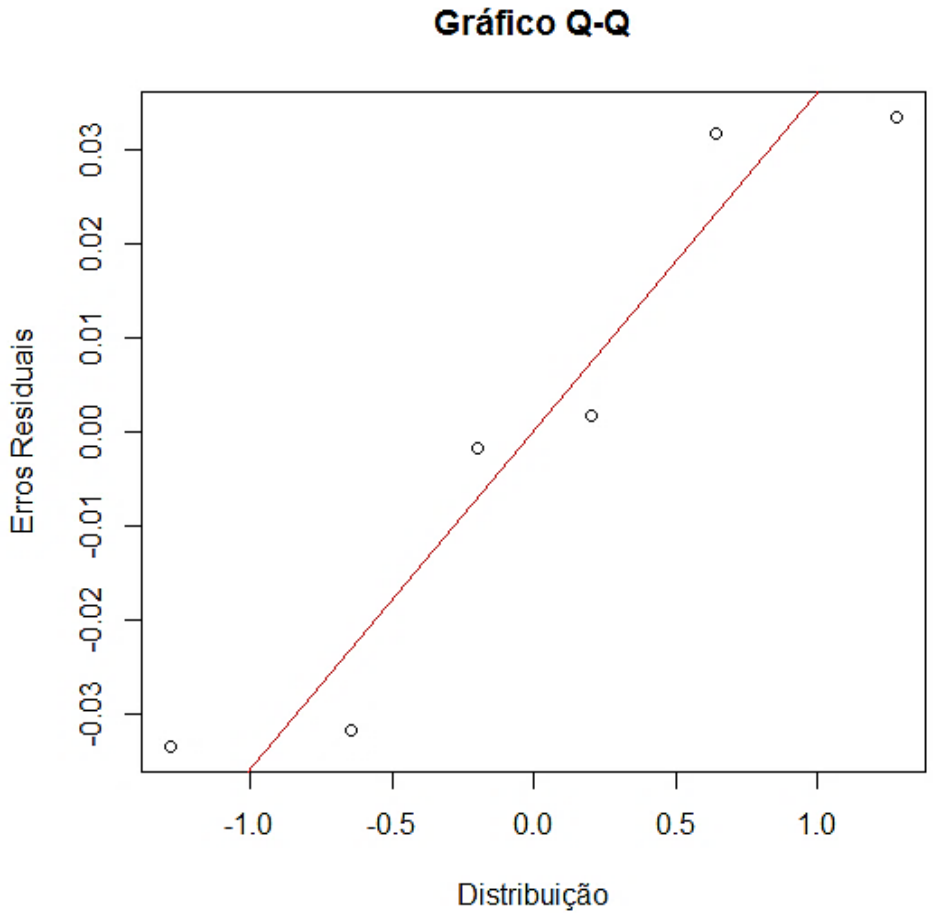
Fonte: Elaborada pelo autor.

Como a análise realizada neste trabalho requer significância de 95%, a tabela F possui um valor de 18,51. Uma vez que o cálculo de F obtido pela análise de variância é menor do que o F da tabela (18,51 > 8), não é possível refutar, com significância de 95%, a hipótese nula H2-0 da questão de pesquisa ***QP2***, para a métrica de *medida-F* referente ao modelo de previsão PPM-HMM. Isto é, não é possível afirmar que, estatisticamente, há diferença em utilizar a base de dados do cenário *TodosTrajetos* comparada com a base do cenário *TrajetosMaiorQueDois*, com respeito à *medida-F*.

Para a verificação de que os valores utilizados no teste ANOVA, referentes à métrica de *medida-F*, são oriundos de uma distribuição normal, foi aplicado o teste de normalidade *Shapiro-Wilk* nos erros experimentais. Da aplicação deste teste, foi possível obter um *valor-p = 0,2563* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

O gráfico Q-Q, apresentado na Figura 22, representa outra abordagem para verificação da normalidade. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Figura 22 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica *Medida-F*, referente ao uso do modelo de previsão PPM-HMM, para a Questão de Pesquisa 2.



Fonte: Elaborada pelo autor.

B.3. Testes com o Modelo de Previsão PPM

Esta seção apresenta os experimentos agrupados pelo modelo de previsão PPM puro, ou seja, sem combinação do PPM com qualquer outra técnica. Para o experimento, o *fator F* possui os níveis referentes ao uso da base de dados, cujos valores podem ser a base *TodosTrajetos* e a base *TrajetosMaiorQueDois*, enquanto o *bloco B* refere-se ao percentual de completude da rota a ser testada (15%, 50% e 85%). As variáveis resposta analisadas pelos testes foram *precisão, cobertura* e *medida-F*.

Como são três variáveis resposta a serem analisadas separadamente, foi criada uma subseção para cada uma delas, e elaborada uma quarta subseção destinada apenas aos comentários dos resultados obtidos.

Resultados para a Variável Resposta de Precisão

A Tabela 55 apresenta os resultados obtidos após a realização do experimento. A primeira coluna da tabela é destinada aos níveis do *bloco B* (exceto as duas últimas linhas), enquanto a segunda coluna apresenta os resultados das médias obtidos para a *precisão* referente aos dois níveis do *fator F*, após cinco repetições. A terceira e a quarta colunas apresentam, respectivamente, as médias obtidas por bloco e o efeito gerado pelo bloco. A penúltima linha da tabela apresenta as médias de cada uma das duas alternativas do fator, enquanto a última linha apresenta o efeito gerado pelas alternativas do *fator F*.

Tabela 55 - Resultados obtidos para a métrica de Precisão – Base de dados x Percentual da rota, para o modelo PPM.

Bloco B	***Fator F* – Base de dados**		**Média bloco**	**Efeito bloco**
	TodosTrajetos	*TrajetosMaiorQueDois*		
15%	0,47	0,62	0,55	-0,07 **(β_1)**
50%	0,54	0,67	0,61	-0,01 **(β_2)**
85%	0,64	0,74	0,69	0,08 **(β_3)**
Média alternativas	**0,55 (α_1)**	**0,68 (α_2)**	**0,61**	
Efeito alternativas	**-0,06**	**0,06**		

Fonte: Elaborada pelo autor.

A Tabela 56 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Já a Tabela 57 apresenta os valores das *Somas dos Quadrados* (*SSY, SS0, SSB, SSA, SSE* e *SST*), além dos valores de MSA e de MSE. Ainda nesta tabela, os percentuais de influência do fator, do bloco e dos erros, bem como o cálculo de F para o experimento realizado e o F crítico (F_c) da tabela F também são apresentados.

Tabela 56 - Erros experimentais para a Precisão, referentes ao preditor PPM.

Bloco B	***Fator F* – Base de dados**	
	TodosTrajetos	*TrajetosMaiorQueDois*
15%	-0,01	0,01
50%	0,00	0,00
85%	0,01	-0,01

Fonte: Elaborada pelo autor.

Tabela 57 - Resultados estatísticos consolidados para a Precisão, referentes ao preditor PPM.

Resultados Estatísticos do Modelo PPM – *Precisão*							
SSY:	2,30	*SS0:*	2,26	*SSE:*	0,0006	*MSA:*	0,0241
SSA:	0,0241	*SSB:*	0,021	*SST:*	0,0459	*MSE:*	0,011
Percentual de Influência do		*Fator:*	52,4%	*Valor De F:*	76	*F_c da tabela F (95%)*	18,51
		Bloco:	46,2%				
		Erro:	1,4%				

Fonte: Elaborada pelo autor.

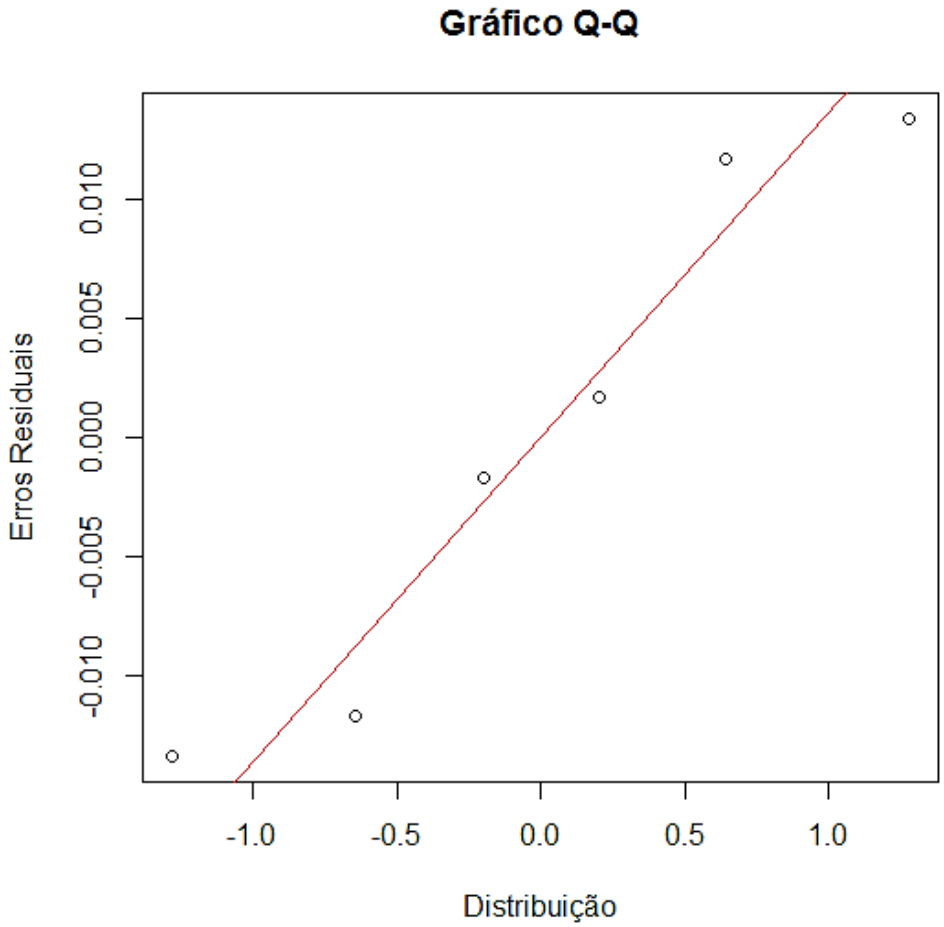
Como a análise realizada neste trabalho requer significância de 95%, a tabela F possui um valor de 18,51. Uma vez que o cálculo de F obtido pela análise de variância é maior do que o F da tabela (18,51 < 76), é possível refutar, com significância de 95%, a hipótese nula H2-0 da questão de pesquisa ***QP2***, para a métrica de *precisão* referente ao modelo de previsão PPM. Isto é, é possível afirmar que, estatisticamente, há diferença em utilizar a base de dados do cenário *TodosTrajetos* comparada com a base do cenário *TrajetosMaiorQueDois*, com respeito à *precisão*.

Com relação aos testes, é importante destacar que a influência do fator e do bloco foram, respectivamente, 52,4% e 46,2%, valores muito próximos. No entanto, conforme o resultado, a tendência é que o uso da base de dados tenha um pouco mais de influência nos resultados obtidos comparado ao bloco utilizado.

Para a verificação de que os valores utilizados no teste ANOVA, referentes à métrica de *precisão*, são oriundos de uma distribuição normal, foi aplicado o teste de normalidade *Shapiro-Wilk* nos erros experimentais. Da aplicação deste teste, foi possível obter um *valor-p = 0,428* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

O gráfico Q-Q, apresentado na Figura 23, representa outra abordagem para verificação da normalidade. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Figura 23 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica *Precisão*, referente ao uso do modelo de previsão PPM, para a Questão de Pesquisa 2.



Fonte: Elaborada pelo autor.

Resultados para a Variável Resposta de Cobertura

A Tabela 58 apresenta os resultados obtidos após a realização do experimento. A primeira coluna da tabela é destinada aos níveis do *bloco B* (exceto as duas últimas linhas), enquanto a segunda coluna apresenta os resultados das médias obtidos para a *cobertura* referente aos dois níveis do *fator F*, após cinco repetições. A terceira e a quarta colunas apresentam, respectivamente, as médias obtidas por bloco e o efeito gerado pelo bloco. A penúltima linha da tabela apresenta as médias de cada uma das duas alternativas do fator, enquanto a última linha apresenta o efeito gerado pelas alternativas do *fator F*.

Tabela 58 - Resultados obtidos para a métrica de Cobertura – Base de dados x Percentual da rota, para o modelo PPM.

Bloco B	***Fator F* – Base de dados**		**Média bloco**	**Efeito bloco**
	TodosTrajetos	*TrajetosMaiorQueDois*		
15%	0,36	0,63	0,50	-0,11 **(β_1)**
50%	0,48	0,73	0,61	0,00 **(β_2)**
85%	0,63	0,79	0,71	0,11 **(β_3)**
Média alternativas	**0,49 (α_1)**	**0,72 (α_2)**	**0,60**	
Efeito alternativas	**-0,11**	**0,11**		

Fonte: Elaborada pelo autor.

A Tabela 59 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Já a Tabela 60 apresenta os valores das *Somas dos Quadrados* (*SSY, SS0, SSB, SSA, SSE* e *SST*), além dos valores de MSA e de MSE. Ainda nesta tabela, os percentuais de influência do fator, do bloco e dos erros, bem como o cálculo de F para o experimento realizado e o F crítico (F_c) da tabela F também são apresentados.

Tabela 59 - Erros experimentais para a Cobertura, referentes ao preditor PPM.

Bloco B	***Fator F* – Base de dados**	
	TodosTrajetos	*TrajetosMaiorQueDois*
15%	-0,02	0,02
50%	-0,01	0,01
85%	0,03	-0,03

Fonte: Elaborada pelo autor.

Tabela 60 - Resultados estatísticos consolidados para a Cobertura, referentes ao preditor PPM.

Resultados Estatísticos do Modelo PPM – *Cobertura*							
SSY:	2,31	*SS0:*	2,18	*SSE:*	0,003	*MSA:*	0,077
SSA:	0,077	*SSB:*	0,046	*SST:*	0,127	*MSE:*	0,002
Percentual de Influência do		*Fator:*	56,7%	*Valor De F:*	44	*F_c da tabela F (95%)*	18,51
		Bloco:	40,1%				
		Erro:	3,2%				

Fonte: Elaborada pelo autor.

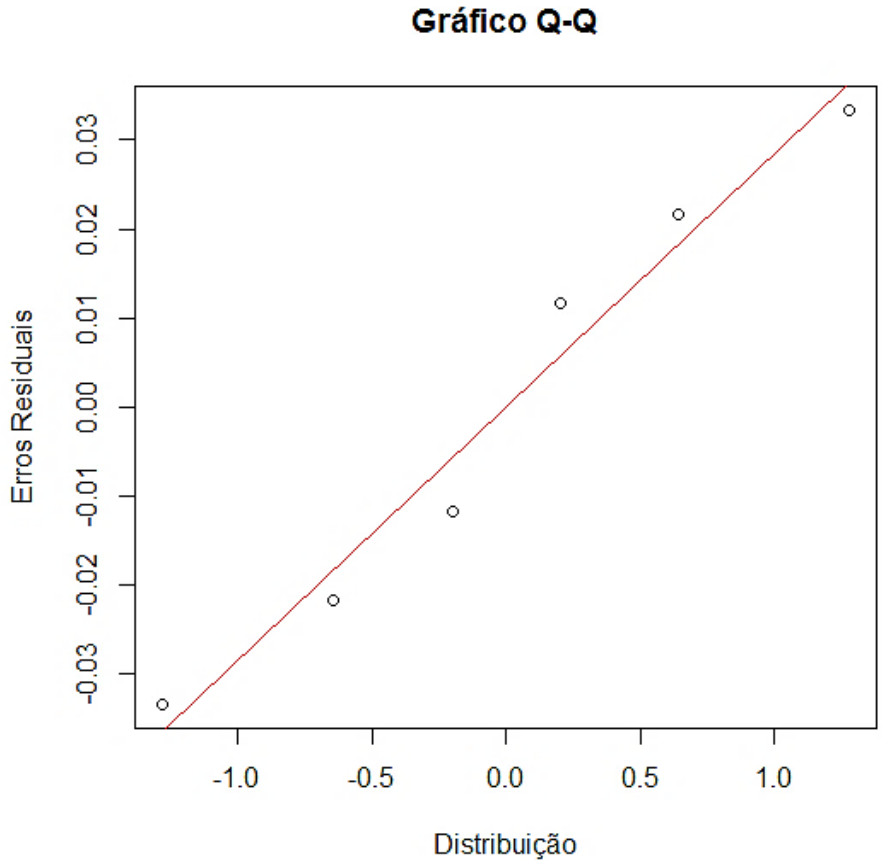
Como a análise realizada neste trabalho requer significância de 95%, a tabela F possui um valor de 18,51. Uma vez que o cálculo de F obtido pela análise de variância é maior do que o F da tabela (18,51 < 44), é possível refutar, com significância de 95%, a

hipótese nula H2-0 da questão de pesquisa ***QP2***, para a métrica de *cobertura* referente ao modelo de previsão PPM. Isto é, é possível afirmar que, estatisticamente, há diferença em utilizar a base de dados do cenário *TodosTrajetos* comparada com a base do cenário *TrajetosMaiorQueDois*, com respeito à *cobertura*.

Para a verificação de que os valores utilizados no teste ANOVA, referentes à métrica de *cobertura*, são oriundos de uma distribuição normal, foi aplicado o teste de normalidade *Shapiro-Wilk* nos erros experimentais. Da aplicação deste teste, foi possível obter um *valor-p = 0,7341* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

O gráfico Q-Q, apresentado na Figura 24, representa outra abordagem para verificação da normalidade. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Figura 24 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica *Cobertura*, referente ao uso do modelo de previsão PPM, para a Questão de Pesquisa 2.



Fonte: Elaborada pelo autor.

Resultados para a Variável Resposta de Medida-F

A Tabela 61 apresenta os resultados obtidos após a realização do experimento. A primeira coluna da tabela é destinada aos níveis do *bloco B* (exceto as duas últimas linhas), enquanto a segunda coluna apresenta os resultados das médias obtidos para a *medida-F* referente aos dois níveis do *fator F*, após cinco repetições. A terceira e a quarta colunas apresentam, respectivamente, as médias obtidas por bloco e o efeito gerado pelo bloco. A penúltima linha da tabela apresenta as médias de cada uma das duas alternativas do fator, enquanto a última linha apresenta o efeito gerado pelas alternativas do *fator F*.

Tabela 61 - Resultados obtidos para a métrica de Medida-F – Base de dados x Percentual da rota, para o modelo PPM.

Bloco B	***Fator F* – Base de dados**		**Média bloco**	**Efeito bloco**
	TodosTrajetos	*TrajetosMaiorQueDois*		
15%	0,41	0,63	0,52	-0,09 **(β_1)**
50%	0,51	0,695	0,60	-0,01 **(β_2)**
85%	0,64	0,76	0,70	0,09 **(β_3)**
Média alternativas	**0,52 (α_1)**	**0,70 (α_2)**	**0,61**	
Efeito alternativas	**-0,09**	**0,09**		

Fonte: Elaborada pelo autor.

A Tabela 62 apresenta os erros experimentais obtidos mediante a aplicação da Equação (1), importante para o cálculo da análise de variância. Já a Tabela 63 apresenta os valores das *Somas dos Quadrados* (*SSY, SS0, SSB, SSA, SSE* e *SST*), além dos valores de MSA e de MSE. Ainda nesta tabela, os percentuais de influência do fator, do bloco e dos erros, bem como o cálculo de F para o experimento realizado e o F crítico (F_c) da tabela F também são apresentados.

Tabela 62 - Erros experimentais para a Medida-F, referentes ao preditor PPM.

Bloco B	***Fator F* – Base de dados**	
	TodosTrajetos	*TrajetosMaiorQueDois*
15%	-0,02	0,02
50%	-0,01	0,01
85%	0,03	-0,03

Fonte: Elaborada pelo autor.

Tabela 63 - Resultados estatísticos consolidados para a Medida-F, referentes ao preditor PPM.

Resultados Estatísticos do Modelo PPM – *Medida-F*							
SSY:	2,30	*SS0:*	2,21	*SSE:*	0,003	*MSA:*	0,05
SSA:	0,05	*SSB:*	0,03	*SST:*	0,08	*MSE:*	0,001
Percentual de Influência do		*Fator:*	77%	*Valor De F:*	35	*F_c da tabela F (95%)*	18,51
		Bloco:	21%				
		Erro:	2%				

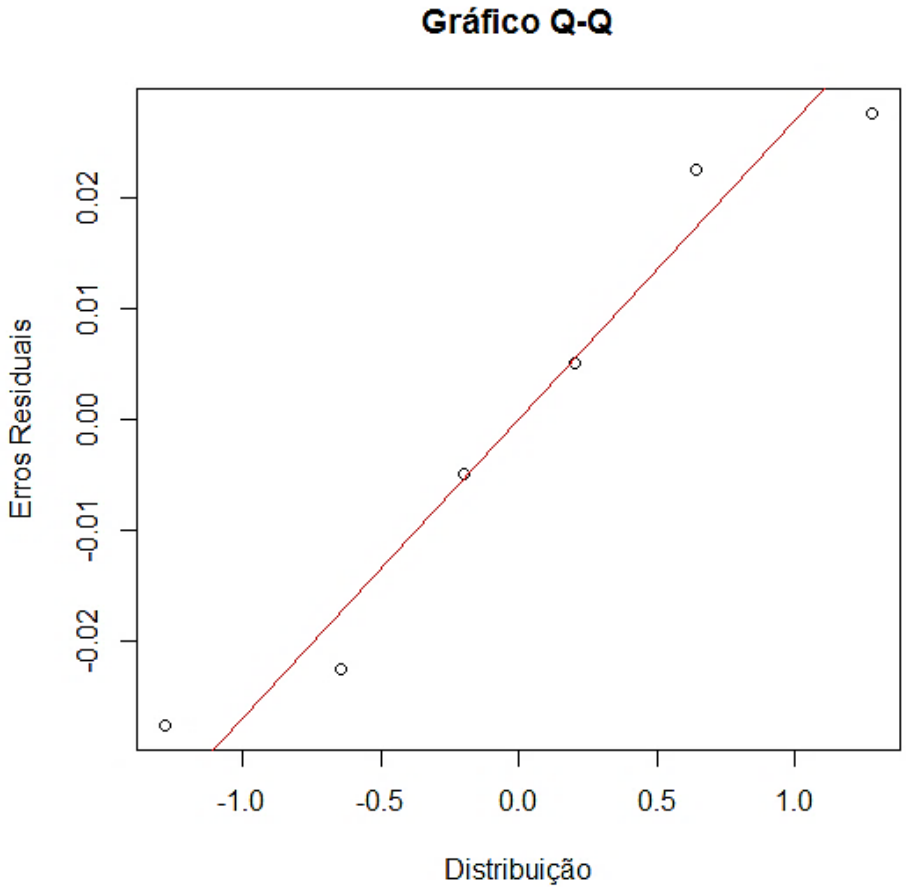
Fonte: Elaborada pelo autor.

Como a análise realizada neste trabalho requer significância de 95%, a tabela F possui um valor de 18,51. Uma vez que o cálculo de F obtido pela análise de variância é maior do que o F da tabela (18,51 < 35), é possível refutar, com significância de 95%, a hipótese nula H2-0 da questão de pesquisa ***QP2***, para a métrica de *medida-F* referente ao modelo de previsão PPM. Isto é, é possível afirmar que, estatisticamente, há diferença em utilizar a base de dados do cenário *TodosTrajetos* comparada com a base do cenário *TrajetosMaiorQueDois*, com respeito à *medida-F*.

Para a verificação de que os valores utilizados no teste ANOVA, referentes à métrica de *medida-F*, são oriundos de uma distribuição normal, foi aplicado o teste de normalidade *Shapiro-Wilk* nos erros experimentais. Da aplicação deste teste, foi possível obter um *valor-p = 0,5631* (> 0,05), resultando, portanto, na afirmação de que não é possível refutar a hipótese nula, isto é, não é possível refutar que os dados utilizados são oriundos de uma distribuição normal, com 95% de nível de confiança.

O gráfico Q-Q, apresentado na Figura 25, representa outra abordagem para verificação da normalidade. Nesta figura, é possível visualizar que a plotagem dos erros residuais não aparenta ter um formato de uma curva nem ao longo dela e nem na extremidade. Além disso, ao traçar uma linha reta próxima aos erros residuais, é possível perceber que esta linha consegue, razoavelmente, representar os erros.

Figura 25 - Gráfico Q-Q que demonstra os erros residuais, obtidos após aplicação do teste ANOVA, para a métrica *Medida-F*, referente ao uso do modelo de previsão PPM, para a Questão de Pesquisa 2.



Fonte: Elaborada pelo autor.